

VU Research Portal

The Restructured Core wordnets in EuroWordNet: Subset1

Vossen, P.J.T.M.; Bloksma, L.

1998

document version

Publisher's PDF, also known as Version of record

[Link to publication in VU Research Portal](#)

citation for published version (APA)

Vossen, P. J. T. M., & Bloksma, L. (1998). *The Restructured Core wordnets in EuroWordNet: Subset1*. Vrije Universiteit.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

vuresearchportal.ub@vu.nl

The Restructured Core wordnets in EuroWordNet: Subset1

Version 3, Final

July 1, 1998

Contributors:

Piek Vossen, Laura Bloksma, University of Amsterdam

Salvador Climent, Maria Antonia Marti, Gabriel Oreggioni, University of Barcelona

Gerard Escudero, German Rigau, Horacio Rodriguez, Universidad Polytechnica de Catalunya

Adriana Roventini, Francesca Bertagna, Antonietta Alonge, Istituto di Linguistica del CNR, Pisa

Carol Peters, Istituto di Elaborazione della Informazione, CNR, Pisa

Wim Peters, University of Sheffield



Deliverable D014, D015, WP3, WP4
EuroWordNet, LE2-4003

Identification number	LE-4003-D014-D015
Type	Document and Lingware
Title	The Restructured Core wordnets in EuroWordNet: Subset1
Status	Final
Deliverable	D-014, D-015
Work Package	WP3 and WP4
Task	T3
Period covered	September 1997 – March 1998
Date	May 22, 1998
Version	2
Number of pages	84
Authors	<ul style="list-style-type: none"> • Piek Vossen, Laura Bloksma, University of Amsterdam • Salvador Climent, Toni Martin, Gabriel <i>Torero</i> Oreggioni University of Barcelona • Gerard Escudero, German Rigau, Horacio Rodriguez, Universidad Polytechnica de Catalunya • Adriana Roventini, Francesca Bertagna, Antonietta Alonge, Istituto di Linguistica del CNR, Pisa • Carol Peters, Istituto di Elaborazione della Informazione, CNR, Pisa • Wim Peters, University of Sheffield
WP/Task responsible	PSA/FUE
Project contact point	Piek Vossen University of Amsterdam Spuistraat 134 1012 VB Amsterdam The Netherlands tel. +31 20 525 4669 fax. +31 20 525 4429 e-mail: Piek.Vossen@hum.uva.nl
EC project officer	Ray Hudson
Status	Public
Actual distribution	Project Consortium, the EuroWordNet User Group, the world via http://www.let.uva.nl/~ewn .
Supplementary notes	n.a.
Key words	Linguistic Resources, Multilingual Wordnets, Language Engineering

Abstract	<p>This deliverable describes the First Subset for Nouns and Verbs in Dutch, Italian, Spanish and English. These First Subsets represent the cores of the wordnets: including the most important meanings on which the other meanings depend. The data are described in terms of tables that specify the synsets, entries, senses and relations, and by comparison with the top ontology distribution and the Parole lexicons. Furthermore, we have carried out two comparisons of the fragments. An in-depth comparison has been carried out for 18 semantic clusters, using the Polaris tool. An overall comparison has been carried out using a graph-matching toolkit developed by FUE. Finally, this deliverable describes the work done for updating the Inter-Lingual-Index (ILI) that interconnects the different wordnets. The conclusions of the overviews and comparisons are being used to guide the final building phase in EuroWordNet.</p>
Status of the abstract	Final
Received on	
Recipient's catalogue number	

Executive Summary

This deliverable describes the First Subset for Nouns and Verbs in Dutch, Italian, Spanish and English. These First Subsets represent the cores of the wordnets: including the most important meanings on which the other meanings depend. The data for each language is described in terms of:

1. quantificational measure: tables that specify the synsets, entries, senses, the language internal relations and the equivalence relations.
2. the top ontology distribution of the synsets, indicating conceptual balancing of the subsets.
3. overlap with the Parole lexicons (as far as available).

Whereas the Spanish wordnet already has reached full coverage (advancing the planning), the Dutch wordnet has just covered the first subset with a higher density of language internal relations, and the Italian wordnet has full coverage but lacks equivalence relations. The distribution of the wordnets over the top-ontology was surprisingly balanced. Some slight imbalances for 1stOrder Entities have to be corrected. Similarly, the overlap with the top-frequent Parole entries is also very high. Missing entries can easily be added.

A better indication on the quality and compatibility is however given by comparing the consistency of the data across the wordnets. For this purpose we have carried out two comparisons of the fragments. An in-depth comparison has been carried out for 18 semantic clusters, using the Polaris tool. An overall comparison has been carried out using a graph-matching toolkit developed by FUE. Both comparisons showed promising results. The in-depth comparison of 18 fields showed reasonable intersections. Most of the mistakes are due to translation errors. Alternative classifications can be used to encode multiple hyperonym. A similar conclusion has been made from the overall comparison. There is a high degree of overlap between subsequences and sequences with 1 gap. By filling these gaps we can improve the coverage in a coordinated way. Furthermore, extremely tangled graphs (Dutch verbs) are mostly due to generation of wrong translations.

The following improvements will therefore be made to the wordnets in the next building phase:

- improve balancing of 1stOrderClusters (Dutch and Italian)
- extend with missing top-frequent Parole entries (Dutch, Italian, Spanish)
- extend coverage (Dutch)
- check translations of extremely long hyponymy chains (especially Dutch verbs)
- fill sequences with 1 gap (Italian, Spanish and Dutch)
- extend translations (Italian)
- improve translation heuristics (Spanish and Dutch)

Finally, this deliverable describes the work done for updating the Inter-Lingual-Index (ILI) that interconnects the different wordnets. The conclusions of the overviews and comparisons are being used to guide the final building phase in EuroWordNet.

Table of Contents

1. General approach for building the wordnets	8
2. Overview results of Subset1	11
2.1 Subset1 for the Dutch wordnet	11
2.2 Subset1 for the Italian wordnet	14
2.3 Subset1 for the Spanish wordnet	16
2.4 Subset1 for the English wordnet	19
2.5 Quantitative conclusions	20
2.6 Overlap with Parole lexicons	22
2.7 Coverage of Subet1 over top concept clusters	24
3. Comparison of the first Subset	28
3.1 Comparing specific semantic fields in the EuroWordNet database	28
3.2 Overall comparison of Subset1	35
3.2.1 Introduction	35
3.2.2 Evaluation of individual wordnets	37
3.2.3 Global evaluation	40
4. Updating the ILI	48
4.1 Clustering Methods	49
4.1.2 (Semi-)automatic clustering	50
4.1.2.1 Sisters	50
4.1.2.2 Autohyponymy	51
4.1.2.3 Twins	51
4.1.2.4 Cousins	51
4.1.2.5 CoreLex	52
4.2 Testing automaticaly created sense groups	52
5 Conclusions	55
References	56
Appendix I In-depth comparison of semantic clusters by different sites	57
Appendix Ia Comparing to the Dutch wordnet	57
Building	58
Comestibles	59
Container	60
Covering	61
Feeling	62
Phenomena	63
Appendix Ib Comparing the Spanish wordnet	64
Garment	65
Furniture	66
Places	67
Plants	68
Sounds	69
Cooking	70
Appendix Ic Comparing the Italian wordnet	71
Animal	71
Artist	72
Worker	73
Instrument	74
Vehicle	75
Movement	76
Knowledge	77
Appendix II Software utilities for graph-comparison	78

List of Tables

Table 1: First Subset Overview NL.....	11
Table 2: Language Internal Relations NL.....	12
Table 3: Equivalence Relations NL	13
Table 4: Status of the Language Internal Relations NL.....	13
Table 5: Status of the Equivalence Relations NL.....	13
Table 6: Reliability of the Euivalence Relations NL.....	13
Table 7: First Subset Overview IT	14
Table 8: Language Internal Relations IT	15
Table 9: Equivalence Relations IT	16
Table 10 : First Subset Overview ES	16
Table 11: Language Internal Relations ES	17
Table 12: Equivalence Relations ES	18
Table 13: Reliability of Equivalence Relations ES	18
Table 14: First Subset Overview GB.....	19
Table 15: Language Internal Relations GB	20
Table 16: Equivalence Relations GB	20
Table 17: First Subset Overview: NL, ES, IT.....	21
Table 18: Overview of Language Internal Relations.....	21
Table 19: Overview of Equivalence Relations.....	22
Table 20: Coverage of Dutch Subset1 related to INL/Celex frequency.....	22
Table 21: Coverage of Spanish Subset1 of Parole Lexicons.....	23
Table 22: Coverage of WordNet1.5 compared to the English Parole Lexicon	23
Table 23: Synsets that are not clustered by the Top Ontology.....	24
Table 24: Nominal Synsets clustered as 1stOrder Concepts.....	24
Table 25: Nominal Synsets clustered as 2ndOrder Concepts.....	26
Table 26: Verbal Synsets clustered as 2ndOrder Concepts	27
Table 27: Nominal Synsets clustered as 3rdOrder Concepts	27
Table 28: Hyperonyms in the wordnets selected for the in-depth comparison	32
Table 29: Projections and Intersections of comparing the First Subset	33
Table 30: ILI chains for nouns.....	37
Table 31: ILI chains for verbs.....	37
Table 32: ILI chains for nouns and verbs	38
Table 33: ILI chains for nouns.....	38
Table 34: clean ILI chains for verbs.....	38
Table 35: Frequencies and ratios of noun chains / length /language	39
Table 36: Frequencies and ratios of verb chains / length /language	39
Table 37: Coverage of noun ILI records.....	40
Table 38: Coverage of verb ILI records	40
Table 39: Coverage of ILI records (total).....	40
Table 40: Coverage of complete noun chains projected over WN1.5 structure.....	41
Table 41: Coverage of complete verb chains projected over WN1.5 structure.....	41
Table 42: Coverage of complete noun chains projected over Dutch wordnet.....	41
Table 43: Coverage of complete verb chains projected over Dutch wordnet	41
Table 44: Coverage of complete noun chains projected over Italian wordnet	41
Table 45: Coverage of complete verb chains projected over Italian wordnet	42
Table 46: Coverage of complete noun chains projected over Spanish wordnet.....	42
Table 47: Coverage of complete verb chains projected over Spanish wordnet.....	42
Table 48: Coverage of partial noun chains of NODES projected over WN1.5 structure.....	42
Table 49: Coverage of partial noun chains of EDGES projected over WN1.5 structure	43
Table 50: Coverage of partial VERB chains of NODES projected over WN1.5 structure	43
Table 51: Coverage of partial VERB chains of EDGES projected over WN1.5 structure	43
Table 52: Coverage of partial noun chains of NODES with 1 gap projected over WN1.5 structure.....	44
Table 53: Coverage of partial noun chains of EDGES with 1 gap projected over WN1.5 structure	44
Table 54: Coverage of partial VERB chains of NODES with 1 gap projected over WN1.5 structure	44
Table 55: Coverage of partial VERB chains of EDGES with 1 gap projected over WN1.5 structure.....	45
Table 56: Coverage of partial noun chains of NODES with 1 gap projected over Dutch wordnet	45

Table 57: Coverage of partial noun chains of EDGES with 1 gap projected over Dutch wordnet.....	45
Table 58: Coverage of partial verb chains of NODES with 1 gap projected over Dutch wordnet	45
Table 59: Coverage of partial verb chains of EDGES with 1 gap projected over Dutch wordnet	45
Table 60: Coverage of partial noun chains of NODES with 1 gap projected over Italian wordnet	45
Table 61: Coverage of partial noun chains of EDGES with 1 gap projected over Italian wordnet	45
Table 62: Coverage of partial VERB chains of NODES with 1 gap projected over Italian wordnet	46
Table 63: Coverage of partial VERB chains of EDGES with 1 gap projected over Italian wordnet.....	46
Table 64: Coverage of partial noun chains of NODES with 1 gap projected over Spanish wordnet	46
Table 65: Coverage of partial noun chains of EDGES with 1 gap projected over Spanish wordnet.....	46
Table 66: Coverage of partial VERB chains of NODES with 1 gap projected over Spanish wordnet.....	46
Table 67: Coverage of partial VERB chains of EDGES with 1 gap projected over Spanish wordnet	46
Table 68: Coverage of partial noun chains of NODES with 2 gaps projected over WN1.5 structure	47
Table 69: Coverage of partial noun chains of EDGES with 2 gaps projected over WN1.5 structure.....	47
Table 70: Automatic derived generalizations and metonymy-relations	53
Table 71: Projection and Intersection increase Dutch-Spanish after adding sense-clusters to the ILI.....	54
Table 72: Errors generated by automatically derived Composite ILIs	54

1. General approach for building the wordnets

The EuroWordNet database is being built (as much as possible) from available existing resources and databases with semantic information developed in various projects. In general, the wordnets are built in two major cycles as indicated by I and II in Figure 1 below. Each cycle consists of a building phase and a comparison phase:

1. Building a wordnet fragment
 - 1.1. Specification of an initial vocabulary
 - 1.2. Encoding of the language-internal relations
 - 1.3. Encoding of the equivalence relations
2. Comparing the wordnet fragments
 - 2.1. Loading of the wordnets in the EuroWordNet database
 - 2.2. Comparing and restructuring the fragments
 - 2.3. Measuring the overlap across the fragments

The building of a fragment is done using local tools and databases which are tailored to the specific nature and possibilities of the available resources. The available resources differ considerably in quality and explicitness of the data. Whereas some sites have the availability of partially structured networks between word senses, others start from genus words extracted from definitions that still have to be disambiguated in meaning.

After the specification of a fragment of the vocabulary, where each site uses similar criteria (there may again be differences due to the different starting points), globally, two approaches are followed for encoding the semantic relations:

Merge model: the selection is done in a local resource and the synsets and their language-internal relations are first developed separately, after which the equivalence relations are generated to WordNet1.5. This approach is followed for the Dutch and Italian wordnets.

Expand model: the selection is done in WordNet1.5 and the WordNet1.5 synsets are translated (using bilingual dictionaries) into equivalent synsets in the other language. The wordnet relations are taken over and where necessary adapted to EuroWordNet. Possibly, monolingual resources are used to verify the wordnet relations imposed on non-English synsets. This approach is followed for the Spanish wordnet.

The Merge model results in a wordnet which is independent of WordNet1.5, possibly maintaining the language-specific properties. The Expand model will result in a wordnet which is very close to WordNet1.5 but which will also be biased by it. Whatever approach is followed also depends on the quality of the available resources.

After a production phase (step Ib and Ic in Figure 1) the results are converted to the EuroWordNet import format and loaded into the common database (step Ic). At that point various consistency checks are carried out, both formally and conceptually. By using the specific options in the EuroWordNet database it is then possible to further inspect and compare the data, to restructure relations where necessary and to measure the overlap in the fragments developed at the separate sites. Those meanings not covered by a site may be included in the extension of the vocabulary in the next building phase.

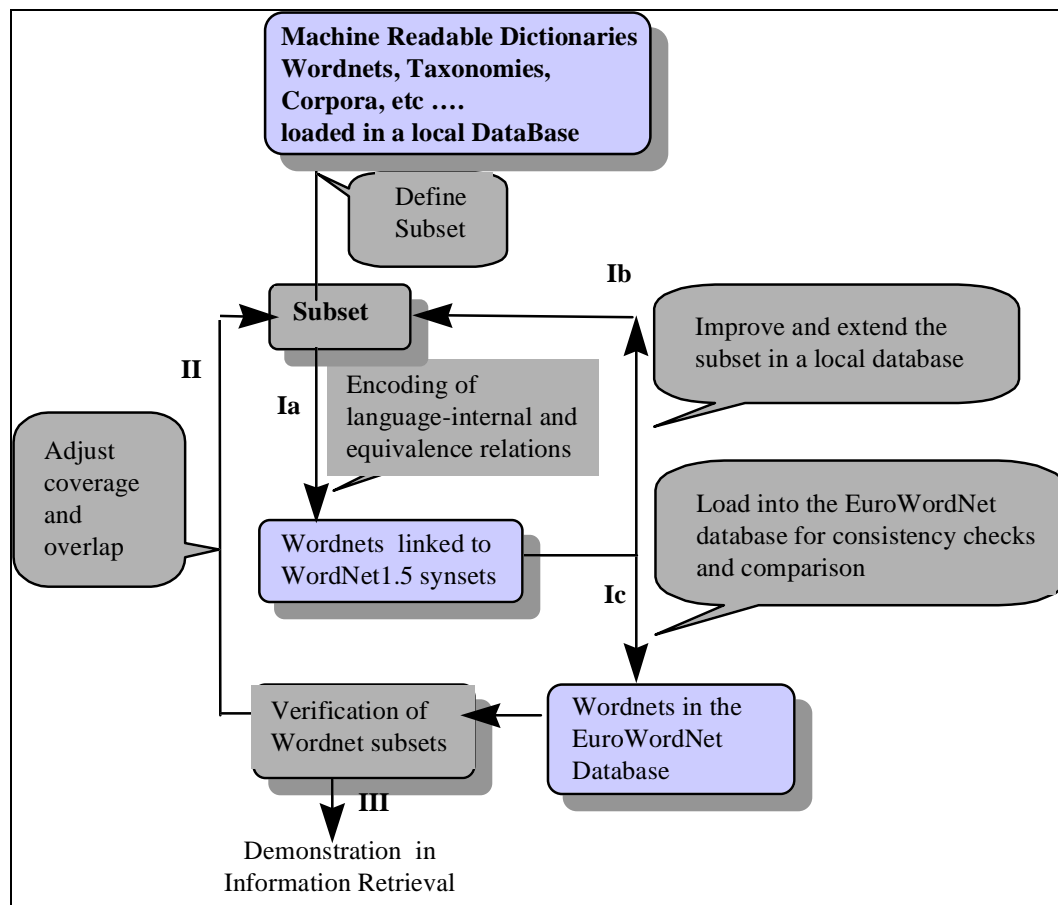


Figure 1.

The overall design of the EuroWordNet database makes it possible to develop the individual language-specific wordnets relatively independently while guaranteeing a minimal level of compatibility. Nevertheless, some specific measures have been taken to enlarge the compatibility of the different resources:

1. The definition of a common set of so-called Base Concepts that is used as a starting point by all the sites to develop the cores of the wordnets. Base Concepts¹ are meanings that play a major role in the wordnets: i.e. have many relations or high positions in the hierarchies.
2. The classification of the Base Concepts in terms of a Top Ontology.
3. The exchange of problems and possible solutions for encoding the relations for the Base Concepts.

The Base Concepts and the Top Ontology are further described in Deliverable D017D034D036 and in [Rodriguez et al. fc.]. In this document we describe the development of the first subset (Subset1) of wordnets in Dutch, Italian, Spanish and English, after the completion of one full cycle. Globally, the building has been carried out starting from the Base Concepts, extending top-down. The general criteria for Subset1 have been:

- All synsets linked to the common set of Base Concepts (1024 synsets).
- All relevant hyperonyms of the synsets related to the Base Concepts.
- The most important hyponyms (1 level down) of the synsets related to the Base Concept

In this way, Subset1 will at least include the core of the different wordnets, including the most important synsets on which more specific meanings depend. The cores will be developed mostly manually, whereas extensions will be derived using semi-automatic techniques.

¹ The notion of Base Concepts should not be confused with Basic-Level Concepts as defined by Rosch (1977). Base Concepts are technically defined as the concepts with most relations. In most cases, they are more general than the Basic Level Concepts.

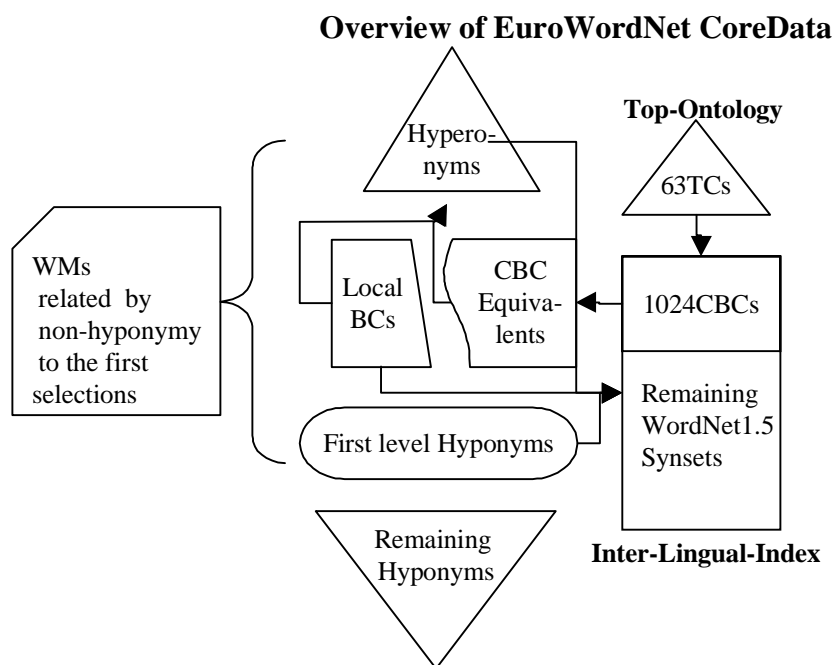
In addition, each site is free to add other concepts, suiting their local approach and starting point. These additions could be:

- synsets related via non-hyponymy relations (such meronymy, role/involvement, antonymy).
- synsets that are translatable to WordNet1.5 synsets.
- Easily extractable from the lexical resources that are available.
- Local Base Concepts, locally important concepts but still not part of the set of common Base Concepts.

The minimal set of synsets aimed at for Subset1 is 10,000 synsets, corresponding with about 20,000 word senses. For each of these synsets the following information has to be minimally specified:

- Hyperonym
- Synonyms (synset members)
- Equivalence relations to WordNet1.5

Optionally, any other relation could be added. The next figure gives an overview of the composition of the vocabulary.



In the case of SHE, a different approach has been followed. Because an English wordnet already exists, SHE has focussed on generating the relations which have been added in EuroWordNet with respect to WordNet1.5. Their selection has been based on the extractability of these relations. SHE has also restructured the ILI by adding missing glosses and grouping senses of words that show some kind of regular polysemy relation. This was necessary to provide a better matching across the wordnets.

The document is then structured in 3 main parts. In section 2, we give for each language overview tables of the covered subset, a comparison of the vocabulary with the Parole² lexicons, and the coverage per Top Ontology clusters (e.g. Communication, Mental, Human, Animal). In section 3, we describe the results of comparing the semantic content of the subsets, by two methodologies: an in-depth comparison of a selection of semantic clusters and a global overall comparison of the complete subsets. Finally, the first results of restructuring the ILI are discussed in section 4. The final Subset1 can be obtained from ELRA (<http://www.icp.inpg.fr/ELRA>) or directly from the builders. Further information on the project and free samples of Subset1 can be downloaded from: <http://www.hum.uva.nl/~ewn>.

² Parole is another EC project that builds lexicons for the most-frequent words with morpho-syntactic information.

2. Overview results of Subset1

The first subset is described in 3 ways for each site:

- number of entries, senses, and synsets covered and the number and kind of relations encoded: sections 2.1, 2.2, 2.3, 2.4., 2.5.
- comparison of the covered vocabulary with Parole lexicons: section 2.6.
- distribution of the vocabulary over the different top-ontology clusters: section 2.7.

2.1 Subset1 for the Dutch wordnet

AMS has followed the Merge Approach, where first a relatively stable Dutch wordnet has been built which has been linked (semi)-automatically to WordNet1.5. The building of the Dutch wordnet has been done by converting the usable relations from the Van Dale database (VLIS) to the EuroWordNet format. Next the converted relations have been verified manually (confirmed or deleted) and missing relations have been added. This manual coding has been done using a special editor, developed at AMS. Since initial hyponymy and synonymy relations were already present in the Van Dale database, we focussed on completing these relations and adding non-hyponymy relations.

The Dutch Subset1 has been based on:

- the common set of Base Concepts
- the local set of Base Concepts
- the hyponyms of the Base Concepts with more than 10 hyponyms themselves
- any other relation which has been manually added or confirmed

AMS has focussed on the encoding of a very solid and stable Subset1 with many different types of Language Internal Relations. We believe that creating a solid and rich semantic context will help both determining the more vague and fuzzy relations such as synonymy, and it will also help determining the equivalence relations with the ILI (WordNet1.5 synsets).

We have manually encoded the equivalence relations for 2,214 synsets. These include the equivalence relations to the common Base Concepts, and the equivalences to the local Base Concepts. All the other equivalences have been generated using a wordnet matching algorithm, partially based on the notion of Conceptual Density as developed by [Agirre and Rigau 1986]. This algorithm weights the senses of translations generated from a bilingual Dutch-English dictionary by comparing the distance of these senses to the senses of the translations of the Dutch semantic context of the word. The context is defined as all word senses that are directly related to it (by means of any semantic relation: hyperonyms, hyponyms, meronyms, etc.). The translation which best fits the translations of the context is selected. By translating synsets partly manually (creating very precise contexts) and by incrementally matching the translation, the best matching translations are generated, gradually improving the context. The current translations are generated after two incremental weightings, and the best 3 translations have been selected for Subset1. An advantage of taking the best 3 equivalence links is that there is a high reliability in coverage (see table below for a sample of Base Concepts). A drawback is that for each correct translation also 2 wrong translations may be generated. An evaluation of the method for more a larger part of the vocabulary will determine what is the best option. On the basis of the comparison it may turn out that we currently generate too much noise on addition to the correct translations.

Table 1: First Subset Overview NL

	<i>Nouns</i>	<i>Verbs</i>	<i>Others</i>	<i>Total</i>
Synsets	5917	3282	389	9588
Number of senses (variants)	10874	5915	1198	17987
X variants per synset	1.84	1.80	3.08	1.88
Corresponding to number of entries (words)	9555	4211	1070	14836
X senses per word	1.14	1.40	1.12	1.21
Language Internal Relations	16917	9486	432	26835
Average per synset	2.86	2.89	1.11	2.80
Equivalent Relations to ILI (WN1.5)	7664	6296	5	13965
Average per synset	1.30	1.92	0.01	1.46
Synset without ILI	1578	394	385	2357

Table 2: Language Internal Relations NL

Language Internal Relations	Nouns	Verbs	Others	Total
Synsets	5917	3282	389	9588
BE_IN_STATE	93			93
CAUSES	140	609		749
HAS_HYPERONYM	6169	3588		9757
HAS_HYPONYM	6169	3588		9757
HAS_HOLONYM	275			275
HAS_HOLO_LOCATION	84			84
HAS_HOLO_MADEOF	97			97
HAS_HOLO_MEMBER	108			108
HAS_HOLO_PART	444			444
HAS_HOLO_PORTION	66			66
HAS_MERONYM	286			286
HAS_MERO_LOCATION	84			84
HAS_MERO_MADEOF	97			97
HAS_MERO_MEMBER	110			110
HAS_MERO_PART	442			442
HAS_MERO_PORTION	65			65
HAS_SUBEVENT	99	109		208
HAS_XPOS_HYPERONYM	9	34	5	48
HAS_XPOS_HYPONYM	34	13	1	48
INVOLVED	54	81		135
INVOLVED_AGENT	4	29		33
INVOLVED_DIRECTION	28	1		29
INVOLVED_INSTRUMENT	5	232		237
INVOLVED_LOCATION	195	21		216
INVOLVED_PATIENT	16	285		301
INVOLVED_SOURCE_DIRECTION	2	1		3
INVOLVED_TARGET_DIRECTION	215	20		235
IS_CAUSED_BY	81	208	320	609
IS_SUBEVENT_OF	91	128		219
NEAR_ANTONYM	132	217		349
NEAR_SYNONYM	138	81		219
ROLE	32			32
ROLE_AGENT	1			1
ROLE_DIRECTION	2			2
ROLE_INSTRUMENT	259			259
ROLE_LOCATION	26			26
ROLE_PATIENT	482			482
ROLE_SOURCE_DIRECTION	16			16
ROLE_TARGET_DIRECTION	18			18
STATE_OF	9	6	79	94
XPOS_NEAR_ANTONYM	3	3		6
XPOS_NEAR_SYNONYM	237	232	27	496
Total	16917	9486	432	26835
Average per synset	2.86	2.89	1.11	2.80

Table 3: Equivalence Relations NL

Equivalence Relations	Nouns	Verbs	Total
EQ_NEAR_SYNONYM	6025	5883	11908
EQ_SYNONYM	1370	375	1745
EQ_HAS_HYPERONYM	174	22	196
EQ_HAS_HYPONYM	85	11	96
EQ_INVOLVED		4	4
EQ_IS_CAUSED_BY	1	1	2
EQ_HAS_HOLONYM	3		3
EQ_HAS_MERONYM	6		6
Total	7664	6296	13960

The next table indicates the number of relations taken over from the Van Dale database or added manually:

Table 4: Status of the Language Internal Relations NL

Language Internal Relations	Nouns	Verbs	Other	Total	Percentages	
Vlis & Okay	2867	2205	3	5075	64.16%	of Vlis Total
Vlis & ?	2317	517	1	2835	35.84%	of Vlis Total
Vlis Total	5184	2722	4	7910	60.40%	of All
Manual & Okay	437	2915	286	3638	70.14%	of manual Total
Manual & ?	1238	258	53	1549	29.86%	of manual Total
Manual Total	1675	3173	339	5187	39.60%	of All
Total	6859	5895	343	13097		

The next table gives the distribution of the manually generated translations and the translations generated by the matching heuristics.

Table 5: Status of the Equivalence Relations NL

Language External relations	Nouns	Verbs	Other	Total	Percentages	
Heuristics & Okay	303	69		372	3.17%	of Heuristics Total
Heuristics & ?	5619	5760		11379	96.83%	of Heuristics Total
Heuristics Total	5922	5829	0	11751	84.15%	of All
Manual & Okay	691	151	5	847	38.26%	of Manual Total
Manual & ?	1051	316		1367	61.74%	of Manual Total
Manual Total	1742	467	5	2214	15.85%	of All
Total	7664	6296	5	13965		

For a sample of Base Concepts we generated the equivalence relations by heuristics and checked the quality of the scores. The next tables gives the reliability of taking the top-3 equivalence relations generated by the heuristics. The table is differentiated for a sample of 1stOrderEntities (FOEs) and 2nd/3rdOrderEntities (HOEs):

Table 6: Reliability of the Euivalence Relations NL

Matching Rank	HOEs		FOEs	
	No of synsets	Perc.	No of Synsets	Perc.
1st score	49	44.95%	40	51.95%
2nd score	36	33.03%	15	19.48%
3rd score	9	8.26%	8	10.39%
>	15	13.76%	14	18.18%
Sum	109		77	

The table shows that in about 50% the 1st score is also the correct translation, and in about 82-87% the correct one is among the top-3. Note that these BCs are the most difficult cases to translate. For more specific concepts the rates will go up.

2.2 Subset1 for the Italian wordnet

At Pisa we have first automatically extracted first level hyponyms of the common Base Concepts (CBCs) from our LDB (which contains data from different sources). As far as the nouns are concerned a sense disambiguation of hyponyms had already been performed within other research projects, thus we only had to revise taxonomies in order to see if hyponyms had been properly assigned to each taxonomy. With respect to the verbs, instead, disambiguated taxonomies had been previously built only for some of our BCs. Thus we had to manually perform a sense disambiguation of most of the taxonomies built. Then, since with respect to other relations to be encoded in EWN our database contained only some information already partially encoded in previous projects (e.g. Acquilex, Delis: synonymy, part-of, set-of, deverbal, deadjectival for nouns; synonymy for verbs) , we had to manually add all the other relations, by analysing mainly definitions, but also other information available (e.g., examples provided for each word sense in our source). The PSA Subset1 has been based on:

- the common set of Base Concepts
- the local set of Base Concepts
- all first level hyponyms of the Base Concepts
- for some taxonomies, also other level hyponyms
- any other relation which has been manually added, by analysing mainly definitions

By using a semi-automatic procedure, part of the data elaborated has been already mapped to WN 1.5., but this work is still in progress. For about 3,091 synsets we have semi-automatically generated the equivalence relations, which have however been verified manually.

Table 7: First Subset Overview IT

	<i>Nouns</i>	<i>Verbs</i>	<i>Others</i>	<i>Total</i>
Synsets	18934	3692	1581	24207
Number of senses (variants)	19646	4577	1587	25810
X variants per synset	1.03	1.24	1	1.09
Corresponding to number of entries (words)	13965	3170		17135
X senses per word	1.40	1.44		1.50
Language Internal Relations	47090	9070		56160
Average per synset	2.48	2.45		2.32
Equivalent Relations to ILI (WN1.5)	5124	653		5777
Average per synset	0.27	0.17		0.22
Synset without ILI	13957	3109	1581	18647

Table 8: Language Internal Relations IT

Language Internal Relations	Nouns	Verbs	Others	Total
Synsets	18934	3692	1581	24207
BE_IN_STATE	123			123
CAUSES		569		569
HAS_HYPERONYM	18654	3651		22305
HAS_HYPONYM	18654	3651		
HAS_HOLONYM				
HAS_HOLO_LOCATION	6			6
HAS_HOLO_MADEOF	1			1
HAS_HOLO_MEMBER	34			34
HAS_HOLO_PART	290			290
HAS_HOLO_PORTION				
HAS_MERONYM	264			264
HAS_MERO_LOCATION	5			5
HAS_MERO_MADEOF	165			165
HAS_MERO_MEMBER	186			186
HAS_MERO_PART	219			219
HAS_MERO_PORTION				
HAS_SUBEVENT		82		82
HAS_XPOS_HYPERONYM	2			2
HAS_XPOS_HYPONYM				
INVOLVED		755		755
INVOLVED_AGENT		36		36
INVOLVED_DIRECTION		5		5
INVOLVED_INSTRUMENT		94		94
INVOLVED_LOCATION	1	10		11
INVOLVED_PATIENT		101		101
INVOLVED_SOURCE_DIRECTION		53		53
INVOLVED_TARGET_DIRECTION		18		18
IS_CAUSED_BY		32		32
IS_SUBEVENT_OF		5		5
NEAR_ANTONYM (ANTONYM)	20	4		24
NEAR_SYNONYM	221	4		225
ROLE	21			21
ROLE_AGENT	1095			1095
ROLE_DIRECTION				
ROLE_INSTRUMENT	80			80
ROLE_LOCATION	51			51
ROLE_PATIENT	16			16
ROLE_SOURCE_DIRECTION				
ROLE_TARGET_DIRECTION				
STATE_OF				
XPOS_NEAR_ANTONYM				
XPOS_NEAR_SYNONYM	7505			7505
Total	47090	9070		56160
Average per synset	2.48	2.45		

Table 9: Equivalence Relations IT

<i>Equivalence Relations</i>	<i>Nouns</i>	<i>Verbs</i>	<i>Total</i>
EQ_SYNONYM	3697	307	4004
EQ_NEAR_SYNONYM	631	259	890
EQ_HAS_HYPERONYM	1947	77	2024
EQ_HAS_HYPONYM		10	10
Total	6275	653	6928

2.3 Subset1 for the Spanish wordnet

FUE has followed the Expand Approach, differentiated for Nouns and Verbs. This approach is based on automatic assignment of Spanish Words to WordNet 1.5 synsets plus further manual revision of the Spanish synsets and relations thus built. The main consequences of this approach are the following:

- A great number of Spanish synsets have been built thus almost reaching already the quantity expected for the end of the project
- The Spanish WordNet (SWN) semantic network is already achieved in terms of the main relations importable from WordNet 1.5 (hypernymy/hyponymy for all Nouns and Verbs, meronymy for Nouns, and causation for Verbs)
- All synsets in the Spanish SWN have an equivalence link to the ILI
- The building of the SWN in this phase has been constrained by conditions of translability to English
- Spanish Nouns automatically assigned to synsets are subject to a degree of confidence, which in any case ranks at least above 85%; all Verbs have been manually checked and corrected; all relations in Subset1 either have been manually built or have a confidence score of 100%.

The building of the SWN Subset1 has proceeded as follows. For Verbs, the PIRAPIDES database (developed by the Universities of Barcelona and Maryland in a join project, see [Dorr et al 1997]) has been used. It consists of 3600 English verb forms organized around Levin's Semantic Classes [Levin 1993], connected to WN1.5 senses, and ambiguously translated to Spanish. It also contains thematic role and diathesis information. Using the latter information and other linguistic knowledge, the database has been manually processed to produce correct SWN synsets. Subset1 includes those which have been already processed; the rest will be included in Subset2. For Nouns, a methodology to map Spanish word forms to WN1.5 synsets using bilingual dictionaries (described in [Atserias et al. 1997]) has been followed. By this procedure, several heuristics have been manually tested using a local lexicological environment in order to choose those which give higher mapping confidence ratios, thus building the appropriate SWN synsets. Furthermore, a number of synsets, including the common set of Base Concepts and the Spanish counterparts of the higher levels in the WN1.5 taxonomy have been manually built. Relations between synsets have been manually checked to include those which are importable from WN1.5 to the SWN. Those which are not will be included in Subset2. Quantity and Quality of the SWN Subset1 can be seen in the tables below.

Table 10 : First Subset Overview ES

	<i>Nouns</i>	<i>Verbs</i>	<i>Others</i>	<i>Total</i>
Synsets	18577	2602	0	21179
number of senses (variants)	41292	6795	0	48087
X variants per synset	2.22	2.61	0	2.27
Corresponding to number of entries (words)	23216	2278	0	25494
X senses per word	1.77	2.98	0	1.88
Language Internal Relations	40559	3749	0	44308
Average per synset	2.18	1.44	0	2.09
Equivalent Relations to ILI (WN1.5)	18634	2602	0	21236
Average per synset	1.00	1.00	0	1.00
Synset without ILI	0	0	0	0
Percentage of Synsets without translation	0%	0%		0%

Table 11: Language Internal Relations ES

Language Internal Relations	Nouns	Verbs	Others	Total
BE_IN_STATE	0	0	0	0
CAUSES	0	40	0	40
HAS_HYPERONYM	18907	1830	0	20737
HAS_HYPONYM	18907	1830	0	20737
HAS_HOLONYM	0	0	0	0
HAS_HOLO_LOCATION	0	0	0	0
HAS_HOLO_MADEOF	75	0	0	75
HAS_HOLO_MEMBER	188	0	0	188
HAS_HOLO_PART	1103	0	0	1103
HAS_HOLO_PORTION	0	0	0	0
HAS_MERONYM	0	0	0	0
HAS_MERO_LOCATION	0	0	0	0
HAS_MERO_MADEOF	75	0	0	75
HAS_MERO_MEMBER	188	0	0	188
HAS_MERO_PART	1103	0	0	1103
HAS_MERO_PORTION	0	0	0	0
HAS_SUBEVENT	0	1	0	1
HAS_XPOS_HYPERONYM	0	0	0	0
HAS_XPOS_HYPONYM	0	0	0	0
INVOLVED	0	3	0	3
INVOLVED_AGENT	0	2	0	2
INVOLVED_DIRECTION	0	0	0	0
INVOLVED_INSTRUMENT	0	1	0	1
INVOLVED_LOCATION	0	0	0	0
INVOLVED_PATIENT	0	0	0	0
INVOLVED_SOURCE_DIRECTION	0	0	0	0
INVOLVED_TARGET_DIRECTION	0	0	0	0
IS_CAUSED_BY	0	40	0	40
IS_SUBEVENT_OF	0	1	0	1
NEAR_ANTONYM	0	0	0	0
NEAR_SYNONYM	6	0	0	6
ROLE	3	0	0	3
ROLE_AGENT	2	0	0	2
ROLE_DIRECTION	0	0	0	0
ROLE_INSTRUMENT	1	0	0	1
ROLE_LOCATION	0	0	0	0
ROLE_PATIENT	0	0	0	0
ROLE_SOURCE_DIRECTION	0	0	0	0
ROLE_TARGET_DIRECTION	0	0	0	0
STATE_OF	0	0	0	0
XPOS_NEAR_ANTONYM	0	0	0	0
XPOS_NEAR_SYNONYM	1	1	0	2
Total	40559	3749	0	44308

Table 12: Equivalence Relations ES

<i>Equivalence Relations</i>	<i>Nouns</i>	<i>Verbs</i>	<i>Total</i>
EQ_NEAR_SYNONYM	0	0	0
EQ_SYNONYM	18577	2602	21179
EQ_HAS_HYPERONYM	40	0	40
EQ_HAS_HYPONYM	14	0	14
EQ_INVOLVED	0	0	0
EQ_IS_CAUSED_BY	0	0	0
EQ_HAS_HOLONYM	1	0	1
EQ_HAS_MERONYM	2	0	2
Total	18634	2602	21236

The next table indicates the reliability of generated translation:

Table 13: Reliability of Equivalence Relations ES

<i>Confidence (Variants)</i>	<i>Nouns</i>	<i>Verbs</i>	<i>Total</i>
100% (Manual)	5041	6795	11836
>97%	403	0	403
>95%	304	0	304
>93%	1598	0	1598
>86%	27649	0	27649
>85%	4625	0	4625
Total	39620	6795	46415

2.4 Subset1 for the English wordnet

Sheffield has concentrated on morphological derivational relations between nouns and verbs in order to create cross-part-of-speech relations expressing morphological as well as semantic links. This subset contains morphological derivational relations between nouns and verbs where the verb has been the base form for the derivational process. The data has been obtained from the CELEX database, in which suffixation and conversion (zero-derivation) processes have been identified and a hierarchical morphological decomposition of the derived form has been performed. CELEX noun-verb pairs with a derivational relation have then been matched against the WordNet wordforms. The WordNet senses of the selected pairs have been manually compared and semantic relations have been manually assigned. The English subset has been based on an extended base concept set of 2277 noun synsets and 567 verb synsets.

Table 14: First Subset Overview GB

	<i>Nouns</i>	<i>Verbs</i>	<i>Others</i>	<i>Total</i>
Synsets	968	894		1862
Number of senses (variants)	2235	3035		5270
X variants per synset	1.99	2.69		2.34
Corresponding to number of entries (words)	1927	2411		4338
X senses per word	1.16	1.26		1.21
Language Internal Relations	2785	2616		5401
Average per synset	2.88	2.92		2.9
Equivalent Relations to ILI (WN1.5)	968	894		1862
Average per synset	1.30	1.92	0.01	1.46
Synset without ILI				

Table 15: Language Internal Relations GB

Language Internal Relations	Nouns	Verbs	Others	Total
Synsets	968	894		1862
CAUSES	24	242		266
HAS_XPOS_HYPERONYM	10	12		22
HAS_XPOS_HYPONYM	12	10		22
HAS_SUBEVENT	6	4		10
XPOS_NEAR_SYNONYM	492	492		984
XPOS_NEAR_ANTONYM	4	4		8
XPOS_FUZZYNYM	169			169
HAS_DERIVED		1399		1399
INVOLVED_AGENT		274		274
INVOLVED_TARGET_DIRECTION		1		1
INVOLVED_INSTRUMENT		104		104
INVOLVED_PATIENT		31		31
INVOLVED		13		13
IS_CAUSED_BY	242	24		266
IS_SUBEVENT_OF	4	6		10
ROLE	13			13
ROLE_AGENT	274			274
ROLE_INSTRUMENT	104			104
ROLE_PATIENT	31			31
ROLE_TARGET_DIRECTION	1			1
DERIVED_FROM	1399			1399
Total	2785	2616		5401
Average per synset	2.87	2.92		2.89

Table 16: Equivalence Relations GB

Equivalence Relations	Nouns	Verbs	Total
EQ_SYNONYM	968	894	1862
Total	968	894	1862

2.5 Quantitative conclusions

The total size of the wordnets aimed at is 25,000 synsets, about 50,000 word senses (synset variants) and 20,000 entries. The next table shows that the figure for the number of synsets has already been reached for Italian and Spanish: 24,207 and 21,179 synsets respectively. The coverage of the Dutch wordnet is much lower (about 50%) but still within the limit which was set for the first subset: 10,000 synsets. This difference only applies to nouns, the verbs are covered equally well in all 3 sites. The main reason for the lower coverage of nouns in the Dutch wordnet is the fact that only those synsets are included that have been processed manually, encoding a maximum of relations. In fact, a much larger Dutch fragment can be provided with hyponymy, synonymy and equivalence relations but this information needs to be verified first. In general, we can thus conclude that the project is advancing the original planning for the first subset. The remaining work will therefore not focus on extending the size of the wordnets but on improving the quality and the overlap across the wordnets (see below).

With respect to the quality, we can already draw some conclusion from table 17. First of all we see that the distribution of senses per synset and per entry is very different for each site. The Spanish synsets contain more variants (double compared to Italian) and also more senses per entry. Since they expanded the WordNet1.5 synsets with Spanish translations, these figures reflect the WordNet1.5 distribution. WordNet1.5 uses a wider notion of synonymy and a more fine-grained differentiation of senses than the traditional dictionaries on which the Italian and Dutch wordnets are based.

Table 17: First Subset Overview: NL, ES, IT³

	Dutch				Italian				Spanish		
	Noun	Verb	Oth.	Total	Noun	Verb	Oth.	Total	Noun	Verb	Total
Synsets	5917	3282	389	9588	18934	3692	1581	24207	18577	2602	21179
Number of senses	10874	5915	1198	17987	19646	4577	1587	25810	41292	6795	48087
Senses per synset	1.84	1.80	3.08	1.88	1.03	1.24	1	1.09	2.22	2.61	2.27
Entries	9555	4211	1070	14836	13965	3170		17135	23216	2278	25494
Senses / entry	1.14	1.40	1.12	1.21	1.40	1.44		1.50	1.77	2.98	1.88
Language Internal Rels.	16917	9486	432	26835	47090	9070		56160	40559	3749	44308
LI Rels./ synset	2.86	2.89	1.11	2.80	2.48	2.45		2.32	2.18	1.44	2.09
Equivalent Rels to ILI	7664	6296	5	13965	5124	653		5777	18634	2602	21236
Eq Rels / synset	1.30	1.92	0.01	1.46	0.27	0.17		0.22	1.00	1.00	1.00
Synsets without ILI	1578	394	385	2357	13957	3109	1581	18647	0	0	0

The differences in language-internal and equivalence relations per synset indicate a further difference in quality. The Dutch wordnet has the highest average of language-internal relations and the Spanish wordnet has the most equivalence relations. In fact, the Spanish equivalence-matching is 1:1 because of the followed procedure. Because they include synsets that can be translated from WordNet1.5, there are no synsets without ILI-references. The next overview tables show more details on these differences.

Table 18: Overview of Language Internal Relations

Language Internal Relations	Dutch		Italian		Spanish		English	
HAS_HYPERONYM	9757	101,76% ⁴	20860	88,67%	20737	97,91%		
HAS_HYPONYM	9757	101,76%	20860	88,67%	20737	97,91%		
HAS_XPOS_HYPERONYM	48	0,50%	2	0,009%			22	1,18%
HAS_XPOS_HYPONYM	48	0,50%					22	1,18%
HAS_HOLONYM	1074	11,20%	331	1,40%	1366	6,45%		
HAS_MERONYM	1084	11,30%	839	3,56%	1366	6,45%		
INVOLVED	1189	12,40%	509	2,16%	6	0,02%	1	0,05%
ROLE	836	8,71%	147	0,62%	6	0,02%	1	0,05%
CAUSES	749	7,81%	468	1,99%	40	0,18%	266	14,28%
IS_CAUSED_BY	609	6,35%	425	1,80%	40	0,18%	266	14,28%
HAS_SUBEVENT	208	2,16%	34	0,14%	1	0,005%	10	0,53%
IS_SUBEVENT_OF	219	2,28%	1	0,004%	1	0,005%	10	0,53%
NEAR_ANTONYM	349	3,64%	20	0,08%				
NEAR_SYNONYM	219	2,28%	225	0,95%	6	0,02%		
BE_IN_STATE	93	0,97%	123	0,52%				
STATE_OF	94	0,98%						
XPOS_NEAR_ANTONYM	6	0,06%					984	52,84%
XPOS_NEAR_SYNONYM	496	5,17%	9082	38,60%	2	0,009%	8	0,43%
HAS_DERIVED							1399	75,13%
Total	26835		57091		44308		5401	
Synsets	9588		23523		21179		1862	
Rels/Synset	2.80		2.42		2.09		2.89	

The first column gives the absolute number of relations per type, the second column for each language gives the relative percentage of the relation for all the covered synsets. Except for English, hyponymy is almost 100% covered. This means that each synset has at least one hyperonym average. For the rest, we see that the Dutch wordnet incorporates far more other relations than the other wordnets. This is in line with the strategy followed for the Dutch

³ In this table we did not include the figure from SHE because their subset is too different to be compared.

⁴ The hyperonym relation is more than 100% because synsets may have multiple hyperonyms.

wordnet, to focus on the rich encoding of the most important concepts rather than a large coverage with shallow information. In the case of **XPOS_NEAR_SYNONYM** we see an extreme number of relations for Italian. The English coverage of relations is very different, since they focus on adding XPOS relations missing in WordNet1.5.

Table 19: Overview of Equivalence Relations

	Dutch			Spanish			Italian		
Equivalence Relations	Nouns	Verbs	Total	Nouns	Verbs	Total	Nouns	Verbs	Total
EQ_SYNONYM	1370	375	1745	18577	2602	21179	3697	307	4004
EQ_NEAR_SYNONYM	6025	5883	11908				631	259	890
EQ_HAS_HYPERONYM	174	22	196	40		40	1947	77	2024
EQ_HAS_HYPONYM	85	11	96	14		14		10	10
EQ_INVOLVED		4	4						
EQ_IS_CAUSED_BY	1	1	2						
EQ_HAS_HOLONYM	3		3	1		1			
EQ_HAS_MERONYM	6		6	2		2			
Total	7664	6296	13960	18634	2602	21236	6275	653	6928

Equivalence relations for most of the Dutch nouns and verbs and most of the Spanish nouns are generated automatically. In the Dutch wordnet, the automatically generated equivalences are always of the type EQ_NEAR_SYNONYM, which explains the high figure. All other equivalences are encoded manually. The equivalences for the Spanish verb are all created manually. Equivalence relations for the Italian synsets are generated semi-automatically but are all manually verified. The best-3 equivalences have been chosen for the Dutch wordnet, which explains the high average of equivalence relations. The main work to be done especially for Dutch and also for Spanish is to improve the quality of the equivalence relations. For the Italian wordnets, the quantity of equivalences has to be increased.

2.6 Overlap with Parole lexicons

The aim of the Parole project is to develop morpho-syntactic lexicons for the most frequent words of the European languages. As such Parole is complementary to EuroWordNet. Future developers should be able to take the generic resources of EuroWordNet and Parole to develop combined lexicons for their NLP applications. It is therefore important to make sure that more or less the same vocabulary is covered in both projects. In both projects the most frequent words should be represented. We therefore compared the lexicons in EuroWordNet with Parole for different corpus frequencies. However, for Dutch and Italian the Parole data are not yet available. For Dutch we therefore used the Celex frequency information, which is based on a 40MLN token corpus.

Table 20: Coverage of Dutch Subset1 related to INL/Celex frequency

	Nouns			Verbs		
Frequency	Celex entries	Celex covered	%coverage	Celex entries	Celex covered	%coverage
1001-	1217	910	74.77%	677	597	88.18%
501-1000	939	449	47.82%	455	315	69.23%
251-500	1408	509	36.15%	637	391	61.38%
101-250	3157	893	28.29%	1176	642	54.59%
51-100	3604	748	20.75%	957	440	45.98%
31-50	3380	565	16.72%	695	271	38.99%
21-30	3016	477	15.82%	495	191	38.59%
11-20	5258	722	13.73%	706	265	37.54%
6-10	4804	550	11.45%	567	212	37.39%
3-5	4713	505	10.72%	377	135	35.81%
2	2338	229	9.79%	346	113	32.66%
0	127	25	19.69%	9	2	22.22%
1	30001	2885	9.62%	1725	499	28.93%
overall	63962	9467	14.80%	8822	4073	46.17%

Table 21: Coverage of Spanish Subset1 of Parole Lexicons

Frequency	Nouns			Verbs		
	parole entries	parole covered	%coverage	parole entries	parole covered	%coverage
1001-	147	143	97.28	110	107	97.27
501-1000	261	246	94.25	139	118	84.89
251-500	462	429	92.86	218	172	78.90
101-250	933	863	92.50	381	257	67.45
51-100	959	863	89.99	374	265	70.86
31-50	892	804	90.13	347	185	53.31
21-30	730	632	86.57	286	141	49.30
11-20	1202	978	81.36	469	175	37.31
6-10	1024	790	77.15	360	129	35.83
3-5	968	665	68.70	254	74	29.13
2	435	257	59.08	123	32	26.02
1	643	334	51.94	131	26	19.85
overall	8656	7004	80.91	3192	1681	52.66

These tables show that there is a very high overlap for the higher frequencies. This is according to our expectation that frequent words are also relatively general and basic and therefore are likely to be occur among the Base Concepts. Each site will individually extract the missing top-frequent words and integrate them in the next building phase.

Finally, the next table shows how the WordNet1.5 matches with the English Parole lexicon, where the frequency information is based on the Cobuild frequency information in CELEX. The coverage for most frequencies is very high. There is a strange deviation for verbs with frequencies 3-5 for which we do not have an explanation.

Table 22: Coverage of WordNet1.5 compared to the English Parole Lexicon

Frequency	Nouns			Verbs		
	parole entries	parole covered	%coverage	parole entries	parole covered	%coverage
1001-	767	698	91	335	333	99.4
501-1000	675	604	89.5	251	250	99.6
251-500	947	863	91.1	366	366	100
101-250	1677	1680	99.8	681	681	100
51-100	1556	1571	99	729	729	100
31-50	1376	1376	100	516	516	100
21-30	1090	1091	99.9	332	332	100
11-20	1525	1529	99.8	392	392	100
6-10	1024	1024	100	160	111	69.3
3-5	650	650	100	52	17	32.7
2	218	218	100	15	15	100
1	157	157	100	18	18	100
0 in CELEX	171	171	100	38	38	100
not in CELEX	676	558	82.5	356	287	80.6
overall	12509	12190	97.4	4241	4085	96.3

2.7 Coverage of Subet1 over top concept clusters

As explained in the introduction, the wordnets are built top-down starting with the Base Concepts. Each site is free to include different lexicalizations patterns when extending the vocabulary from the Base Concepts down. To still get an idea of the conceptual distribution of this extension we also measure the progress of the wordnets relative to the Top Ontology, which represents the diversity of Base Concepts that have been selected. For this purpose, AMS implemented an inheritance mechanism that derives the Top Concepts from hyperonyms in WordNet1.5. By loading ILI-equivalences of the Spanish, Dutch and Italian first subset in the Amsterdam lexical database (ALS), it is possible to collect the Top Concepts that apply to these equivalences via hyponymy-inheritance in WordNet1.5. By applying this to all the equivalences, it is possible to quantify the coverage per top concept. Note that this measurement depends on the quality and quantity of the equivalence relations. Not all synsets in the Italian and Dutch wordnets have a (correct) equivalent relation. Furthermore, it may be that the hyponymy relations in the local wordnets are different, but the global semantic classification still has to be consistent. This method therefore still gives a good indication of the conceptual coverage.

The Top Ontology is divided in 3 main parts:

1stOrderEntities (nouns): concrete things

2ndOrderEntities (nouns, verbs and adjectives): states, events, processes, relations and properties

3rdOrderEntities (nouns): idea, knowledge, propositions

However, there are cases where the hyponymy links do not provide any top-concept: i.e. not all WordNet1.5 tops have been classified.

Table 23: Synsets that are not clustered by the Top Ontology

VOID	WN	NL	ES	IT
nouns	0	17	33	15
verbs	2109	310	638	385

WordNet1.5 only has 11 tops for Nouns but 573 for verbs. Most of the noun tops have at least one Top Concept assigned, whereas only 48 of the verb tops have been classified so far. This explains that only a few nominal synsets have not inherited an top concept, whereas a large proportion of the verbs is not (in)directly linked to the ontology. In the near future we will classify all the WordNet1.5 tops so that a complete clustering can be made.

Table 24: Nominal Synsets clustered as 1stOrder Concepts

Nouns	WN		NL			ES			IT		
1stOrderEntity	5	0,00	0	0,00	0,00	0	0,00	0,00	0	0,00	0,00
Animal	4024	2,76	55	0,59	0,04	729	1,68	0,50	577	3,76	0,40
Artifact	12054	8,27	1198	12,86	0,82	4354	10,04	2,99	1137	7,42	0,78
Building	589	0,40	105	1,13	0,07	282	0,65	0,19	14	0,09	0,01
Comestible	2207	1,51	154	1,65	0,11	551	1,27	0,38	132	0,86	0,09
Container	1060	0,73	59	0,63	0,04	321	0,74	0,22	80	0,52	0,05
Covering	1279	0,88	103	1,11	0,07	520	1,20	0,36	21	0,14	0,01
Creature	473	0,32	2	0,02	0,00	50	0,12	0,03	3	0,02	0,00
Function	10183	6,99	538	5,78	0,37	4028	9,29	2,76	1356	8,85	0,93
Functional	120	0,08	18	0,19	0,01	62	0,14	0,04	8	0,05	0,01
Furniture	196	0,13	17	0,18	0,01	68	0,16	0,05	7	0,05	0,00
Garment	446	0,31	22	0,24	0,02	195	0,45	0,13	4	0,03	0,00
Gas	56	0,04	8	0,09	0,01	26	0,06	0,02	27	0,18	0,02
Group	13092	8,98	225	2,42	0,15	1430	3,30	0,98	821	5,36	0,56
Human	6315	4,33	215	2,31	0,15	2741	6,32	1,88	1227	8,01	0,84
ImageRepresentati on	480	0,33	28	0,30	0,02	171	0,39	0,12	15	0,10	0,01
Instrument	4557	3,13	512	5,50	0,35	1676	3,87	1,15	877	5,72	0,60

LanguageRepresentation	1883	1,29	107	1,15	0,07	527	1,22	0,36	32	0,21	0,02
Liquid	1083	0,74	67	0,72	0,05	229	0,53	0,16	91	0,59	0,06
Living	16375	11,23	484	5,20	0,33	4185	9,65	2,87	2455	16,02	1,68
MoneyRepresentation	241	0,17	23	0,25	0,02	81	0,19	0,06	11	0,07	0,01
Natural	15182	10,41	1646	17,67	1,13	5328	12,29	3,66	1312	8,56	0,90
Object	26174	17,96	1864	20,01	1,28	8679	20,02	5,95	3329	21,72	2,28
Occupation	1222	0,84	42	0,45	0,03	571	1,32	0,39	345	2,25	0,24
Part	7412	5,08	587	6,30	0,40	1957	4,51	1,34	183	1,19	0,13
Place	3235	2,22	220	2,36	0,15	856	1,97	0,59	84	0,55	0,06
Plant	5619	3,85	81	0,87	0,06	670	1,55	0,46	674	4,40	0,46
Representation	592	0,41	55	0,59	0,04	245	0,57	0,17	24	0,16	0,02
Software	134	0,09	5	0,05	0,00	29	0,07	0,02	6	0,04	0,00
Solid	3985	2,73	324	3,48	0,22	1157	2,67	0,79	92	0,60	0,06
Substance	5045	3,46	514	5,52	0,35	1451	3,35	1,00	242	1,58	0,17
Vehicle	453	0,31	38	0,41	0,03	189	0,44	0,13	141	0,92	0,10
Total	145771		9316		6,39	43358		29,74	15327		10,51

The first column gives the full list of the 1stOrder Top Concepts. The first column of each wordnet gives the number of synsets (represented as ILI-records) that are either directly or indirectly via a hyperonym chain classified by the Top-Concept. The next column gives the percentage of the total set of 1stOrder nouns covered by each wordnet and the third column for NL, ES and IT gives the percentage of the total set in WordNet1.5. The second columns of each wordnet gives the distribution per top-concept class. If the wordnets are equally balanced then the relative percentages of the wordnets should be the same, even if the total size of the wordnets are different. When a particular percentage is significantly lower than the other wordnets it means that this wordnet should be extended in this domain to become more balanced.⁵

⁵ The table is also useful for users of the wordnets to verify if particular domains or fields of their interest are well-represented or need to be extended.

In this table we clearly see that the Spanish wordnet closely follows the balancing of WordNet1.5, due to the methodology that has been applied. The Dutch and Italian wordnets show a diverging distribution. The fields with relatively lower coverage are marked in the table:

Spanish wordnet: Group.
 Dutch wordnet: Animal; Creature; Group; Human; Living; Occupation; Plant.
 Italian wordnet: Building; Container; Covering; Furniture; Garment; LanguageRepresentation; Part; Place; Solid; Substance.

The differences in distribution do not necessarily imply that the areas are **badly** represented. They can also be due to differences in lexicalization across the languages or to a lack of equivalence relations in a particular area. Nevertheless, each wordnet builder has to check these fields in their resources to find whether these differences are due to incompleteness or due to lexicalization differences. In the former case, the wordnets have to be extended. Note that irregardless of the balancing or distribution of the synsets, the total coverage is much lower than WordNet1.5 and should especially be increased for the Dutch wordnet.

The next two tables shows the distribution for nouns and verbs that are classified as 2ndOrderEntities according to the WordNet1.5 hyponymy chains. Whereas the previous table showed some differences in conceptual coverage, the next tables are remarkably balanced. Only Quantity and Usage are slightly under-represented in the Dutch wordnet and latter in the Italian wordnet.

Table 25: Nominal Synsets clustered as 2ndOrder Concepts

Nouns	WN1.5		NL			ES			IT		
Agentive	6146	6,96	324	5,72	0,37	2539	7,24	2,88	185	6,30	0,21
BoundedEvent	4753	5,38	292	5,16	0,33	1934	5,51	2,19	149	5,08	0,17
Cause	5496	6,23	314	5,55	0,36	2305	6,57	2,61	179	6,10	0,20
Communication	3852	4,36	223	3,94	0,25	1372	3,91	1,55	88	3,00	0,10
Condition	2325	2,63	311	5,49	0,35	994	2,83	1,13	78	2,66	0,09
Dynamic	9400	10,65	610	10,78	0,69	4057	11,56	4,60	377	12,84	0,43
Existence	198	0,22	19	0,34	0,02	97	0,28	0,11	6	0,20	0,01
Experience	4012	4,55	294	5,19	0,33	1754	5,00	1,99	199	6,78	0,23
Location	851	0,96	60	1,06	0,07	343	0,98	0,39	31	1,06	0,04
Manner	573	0,65	29	0,51	0,03	241	0,69	0,27	15	0,51	0,02
Mental	6166	6,99	390	6,89	0,44	2439	6,95	2,76	200	6,81	0,23
Modal	291	0,33	17	0,30	0,02	130	0,37	0,15	18	0,61	0,02
Phenomenal	1216	1,38	144	2,54	0,16	507	1,45	0,57	75	2,56	0,08
Physical	4712	5,34	445	7,86	0,50	1914	5,46	2,17	206	7,02	0,23
Possession	95	0,11	6	0,11	0,01	46	0,13	0,05	1	0,03	0,00
Property	6975	7,90	505	8,92	0,57	3166	9,02	3,59	230	7,84	0,26
Purpose	9250	10,48	459	8,11	0,52	3374	9,62	3,82	237	8,07	0,27
Quantity	2129	2,41	95	1,68	0,11	662	1,89	0,75	82	2,79	0,09
Relation	4148	4,70	248	4,38	0,28	1471	4,19	1,67	97	3,30	0,11
Social	7449	8,44	353	6,24	0,40	2650	7,55	3,00	185	6,30	0,21
Static	3408	3,86	310	5,48	0,35	1201	3,42	1,36	133	4,53	0,15
Stimulating	599	0,68	42	0,74	0,05	298	0,85	0,34	24	0,82	0,03
Time	712	0,81	32	0,57	0,04	227	0,65	0,26	24	0,82	0,03
UnboundedEvent	2792	3,16	124	2,19	0,14	1226	3,49	1,39	113	3,85	0,13
Usage	723	0,82	14	0,25	0,02	138	0,39	0,16	3	0,10	0,00
Total	88271		5660		6,41	35085		39,75	2935		3,32

Table 26: Verbal Synsets clustered as 2ndOrder Concepts

Verbs	WN		NL			ES			IT		
Agentive	3139	9,15	344	7,86	1,00	713	7,52	2,08	289	7,32	0,84
BoundedEvent	4038	11,77	496	11,33	1,45	1084	11,44	3,16	363	9,19	1,06
Cause	3265	9,52	422	9,64	1,23	911	9,61	2,66	371	9,39	1,08
Communication	1067	3,11	123	2,81	0,36	241	2,54	0,70	100	2,53	0,29
Condition	578	1,69	65	1,48	0,19	165	1,74	0,48	77	1,95	0,22
Dynamic	5856	17,07	892	20,37	2,60	1712	18,06	4,99	818	20,71	2,39
Existence	831	2,42	89	2,03	0,26	213	2,25	0,62	86	2,18	0,25
Experience	384	1,12	46	1,05	0,13	143	1,51	0,42	60	1,52	0,17
Location	2579	7,52	408	9,32	1,19	859	9,06	2,50	318	8,05	0,93
Manner	174	0,51	21	0,48	0,06	52	0,55	0,15	15	0,38	0,04
Mental	840	2,45	75	1,71	0,22	224	2,36	0,65	111	2,81	0,32
Modal	16	0,05	2	0,05	0,01	6	0,06	0,02	2	0,05	0,01
Phenomenal	10	0,03	3	0,07	0,01	8	0,08	0,02	0	0,00	0,00
Physical	2938	8,57	328	7,49	0,96	927	9,78	2,70	262	6,63	0,76
Possession	655	1,91	88	2,01	0,26	153	1,61	0,45	54	1,37	0,16
Property	170	0,50	14	0,32	0,04	40	0,42	0,12	22	0,56	0,06
Purpose	1896	5,53	189	4,32	0,55	387	4,08	1,13	139	3,52	0,41
Quantity	302	0,88	40	0,91	0,12	93	0,98	0,27	25	0,63	0,07
Relation	307	0,90	36	0,82	0,10	76	0,80	0,22	38	0,96	0,11
SituationType	73	0,21	14	0,32	0,04	28	0,30	0,08	24	0,61	0,07
Social	1391	4,06	157	3,59	0,46	303	3,20	0,88	144	3,65	0,42
Static	165	0,48	24	0,55	0,07	48	0,51	0,14	32	0,81	0,09
Stimulating	334	0,97	24	0,55	0,07	157	1,66	0,46	95	2,41	0,28
Time	14	0,04	0	0,00	0,00	6	0,06	0,02	1	0,03	0,00
UnboundedEvent	1037	3,02	145	3,31	0,42	256	2,70	0,75	112	2,84	0,33
Usage	129	0,38	24	0,55	0,07	34	0,36	0,10	6	0,15	0,02
Total	34297		4379		12,77	9477		27,63	3949		11,51

The fact that the 2ndOrderEntities are equally balanced may also indicate that the Top-Ontology classification is more shallow compared to the nominal classification. A shallow, more abstract classification necessarily tends to blur out differences as well. From these tables, we nevertheless cannot derive any conclusions for extending the wordnets in a particular direction.

Finally, the next table gives the nominal synsets classified as 3rdOrderEntities, where the percentage give the proportion of the set in WordNet1.5. Here we see that the coverage for Spanish and Dutch is similar to the total coverage of 1stOrder and 2ndOrder Entities compared to WordNet1.5. The Italian wordnet however shows a significantly lower percentage here. This can again either be due to lexicalization differences, incompleteness or to a lack of equivalence relations in this area.

Table 27: Nominal Synsets clustered as 3rdOrder Concepts

	WN	NL	ES	IT			
3rdOrderEntity	4989	309	6,19%	1860	37,28%	147	2,95%

3. Comparison of the first Subset

We have carried out two different types of comparisons for Subset1:

- an in-depth comparison of the wordnets in the EuroWordNet database for 18 semantic fields
- an overall comparison of the full subset

The in-depth-comparison is carried out using the comparison options in the Polaris tool. The overall comparison is done by generating the hyperonym chains for the full subset in the form of the ILI-records. This resulted in compatible graph-structures for each wordnet. FUE has developed a special toolkit for comparing these graph-structures.

3.1 Comparing specific semantic fields in the EuroWordNet database

The goal of the comparison is to measure the quality and quantity of the local wordnet by comparing them to the other wordnets (see [Peters et al., fc] for further details). For the comparison, we make a distinction between the Source wordnet and the Reference wordnets. The Source wordnet is the wordnet at a local site which is going to be evaluated by comparing it to the Reference wordnets. It is not the purpose to evaluate the Reference wordnets. A comparison will give information on:

1. the quality and quantity of equivalence relations
2. the overlap across wordnets
3. the coherence of classification

The following Clusters have been examined for the Subset1 of nouns and verbs, where a division is made between 1stOrderEntities (FOEs) and 2nd/3rdOrderEntities (HighOrderEntities or HOEs):

AMS	FOE	Building, Comestible, Container, Covering
	HOE	Feelings, Phenomena
FUE	FOE	Garment, Place, Furniture, Plant
	HOE	Cooking, Sounds
PSA	FOE	Animal, Human, Instrument, Vehicle
	HOE	Movements, Knowledge

Each site will distribute the major hyperonyms that represent the most important tops of these semantic fields, e.g.: {construction-4} in WordNet1.5, {bouwwerk-1} in Dutch, {construzione-1} in Italian and {construcción-4} in Spanish. The comparison then globally consists of:

- Extract the hyponyms of the Representative hyperonyms in these fields in each wordnet.
- Project the hyponyms of the Reference wordnets to the Source wordnet.
- Compare the projected hyponyms with the hyponyms in the Source wordnet

The projections in the EuroWordNet database result in sets of word meanings (WMs) in the source wordnet related to the same Inter-Lingual-Index concepts. The next screen-dump of the EuroWordNet database (Polaris) shows such a projection from the hyponyms of *construcción-4* in the Spanish wordnet (the left window) to the Dutch wordnet. In the right window the Dutch WMs are shown that are related to the same ILIs given as equivalences of the Spanish hyponyms. The bottom window shows the corresponding ILI-records.

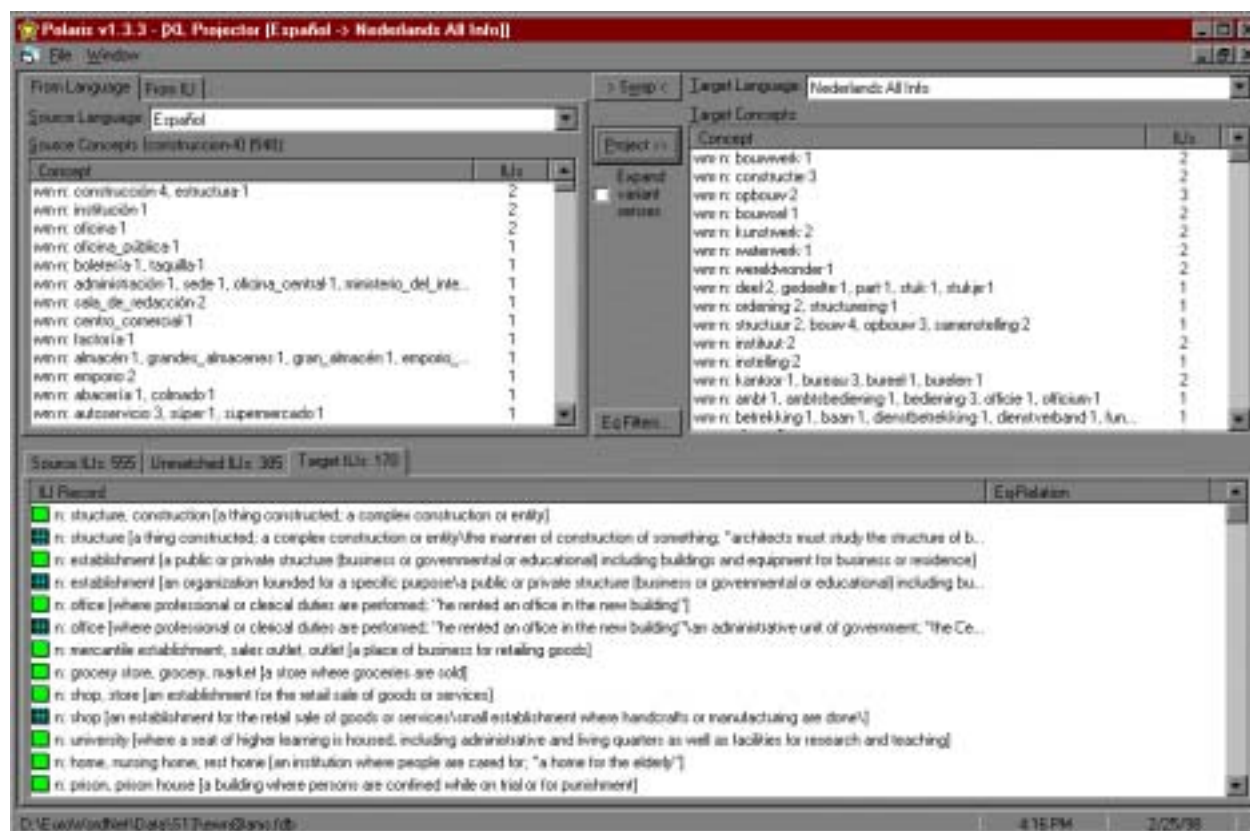


Figure 3: Projection of hyponyms of *construcción-4* in the Spanish wordnet to the Dutch wordnet.

A projection may partly overlap with the set of hyponyms build up in the local Source wordnet (the hyponyms of *bouwwerk-1* in Dutch). It is possible to compare sets of WMs in the database and to derive the intersection, union and difference. The intersection represents the degree of compatibility of wordnets. The WMs which are unique in the projection of the Reference wordnets or which are unique in the Source wordnet hyponyms represent the incompatibility of the wordnets. Unique sets of WMs are thus both present in the Source wordnet (i.e. the ILI-records generated by the projection have also been used in the Source-wordnet to link local synsets to the ILI) but are classified differently. The projection which is unique in the Reference wordnet is then apparently not a hyponym (at any level) of “*bouwwerk 1*” (construction 4) in the Dutch wordnet. On the other hand, the ILIs linked to the WMs which are unique in the Source wordnet are not part of the set of ILIs projected by the Reference wordnets (they may be present but classified differently). Projected WMs which are Unique in the Reference wordnets can be diagnosed as follows:

- **projected by a wrong translation in the Source wordnet:** e.g. “*afsluiting 2*” (the event of blocking a passage or container) is automatically but wrongly translated to an object or construction with that function “*barrier 1*”. The event will not show up as a hyponym of the Dutch equivalent “*bouwwerk*” (construction 4) but it will be generated by the projection of hyponyms of “*construction 4*”.⁶
- **Wrongly classified in the Source wordnet:** e.g. “*onderdak 1*” (shelter 1, place to stay) is classified as “*gelegenheid 1*” (occasion) which is the wrong sense. It should have been classified as “*gelegenheid 2*” (place or building with a purpose).
- **Alternatively classified in the Source wordnet:** e.g. “*centrum 3*” (center 4) is defined in Dutch as “place, institution, building, area, where certain activities take place”. It is only classified in the Dutch wordnet as “institution” and not as “building”.

A more systematic overview can be provided by generating all the hyperonyms for the WMs projected by the Reference wordnet and not included in the Source hyponym set. Below is a list of the most frequent hyperonyms in Dutch for this set:

11 **eigenschap 1 (property)**

12 **organisatie 3 (organisation)**

⁶ This will also show up when we project the Dutch hyponyms back to Dutch. In that case, WMs that are linked to these ILI-records by mistake will also be projected, but they are not part of the original set (if the hyponymy relations are correct).

```

14 woning 1 (living, home)
16 groep 4 (group)
21 steun 1 (support)
27 ruimte 3 (space)
42 plaats 1 (place)
53 deel 2 (part)
70 voorwerp 1 (object)
70 zaak 1 (thing)
71 entiteit 1 (thing)
71 object 1 (object)
430 iets 1 LEAF (anything)

```

The final “iets” (anything) is meaningless, because it is the top of all, but there are also some meaningful hyperonyms. The cluster “woning” (living), “ruimte” (space), “plaats” (place) represents *places* not classified as *constructions* but many of these may very well get an additional hyponymy link to construction. Those linked to “organisatie” (organisation) will either be solved by so-called EQ_METONYM links to the new sense-groups (after the ILI has been extended with these global senses, see section 4) or they need an extra classification or sense for the construction/building where the institutes are settled. Note that some of these can also have wrong ILI-links. Whenever there are two senses in the Dutch wordnet, one for the *building* and one for the *institute*, they should not be translated to the same sense in the ILI.

The hyperonyms “voorwerp”, “object”, “zaak” represent *physical objects* which are not incompatible with *constructions*, but which are also not very meaningful because they represent a very large and diverse group. Something similar can be said for “deel 2” (part) and “groep 4” (group), which often refer to *parts* or *groups* of *constructions*. It may be the case, that these can still get an additional link to *construction* as well. Finally, “eigenschap” (property) is a hyperonym that is totally incompatible with *constructions*. These must all be errors in the translation or in the classification. The above overview cannot be generated directly by Polaris. It has to be done by either exporting the WMs, which are unique in the Reference wordnets, or by loading the senses in a local tool (which has been done here for the Dutch WMs, using the AMS LDB).

The evaluation of the differences minimally consist of a manual inspection in Polaris by looking at WMs and counting the incorrect cases: i.e. WMs that cannot possibly be *constructions*. Inspection of the WMs that are Unique in the Source wordnets will also be done by hand, going through the list in Polaris. It may be the case that these WMs are not covered in the Reference wordnets, or that one of the above explanations applies. However, we will not evaluate the Reference wordnets, so it suffices to browse through the list and count the number of WMs that do not belong there.

Finally, in some cases the projection of a Reference wordnet may not generated output in the Source wordnet. The unmatched ILIs can be projected by taking all the senses of the variants. This generates a lot of WMs in the Source wordnet which fall outside the scope of the comparison. However, after taking the intersection of the projection on a word level with the hyponyms in the Source wordnet, it is possible to filter out near-matches. We speak of near-matches when two synsets are related to the same ILI-word but to different senses. Because of the sense-differentiation in WordNet1.5 this is a likely cause of mismatch across wordnets. The next figure gives an overview of the sets of word meanings that can be generated for hyponyms of *constructions*, where Dutch is the Source wordnet and the other languages represent the Reference wordnets. The remaining cases of WMS that cannot be projected fall into the following classes:

- Not included in the subset of the Source wordnet
- Gaps in the language
- Gaps in the resource

Comparison of hyponyms of “construction” across wordnets

Source Wordnet = NL-net

Reference wordnets = WN1.5, IT-Net, ES-Net

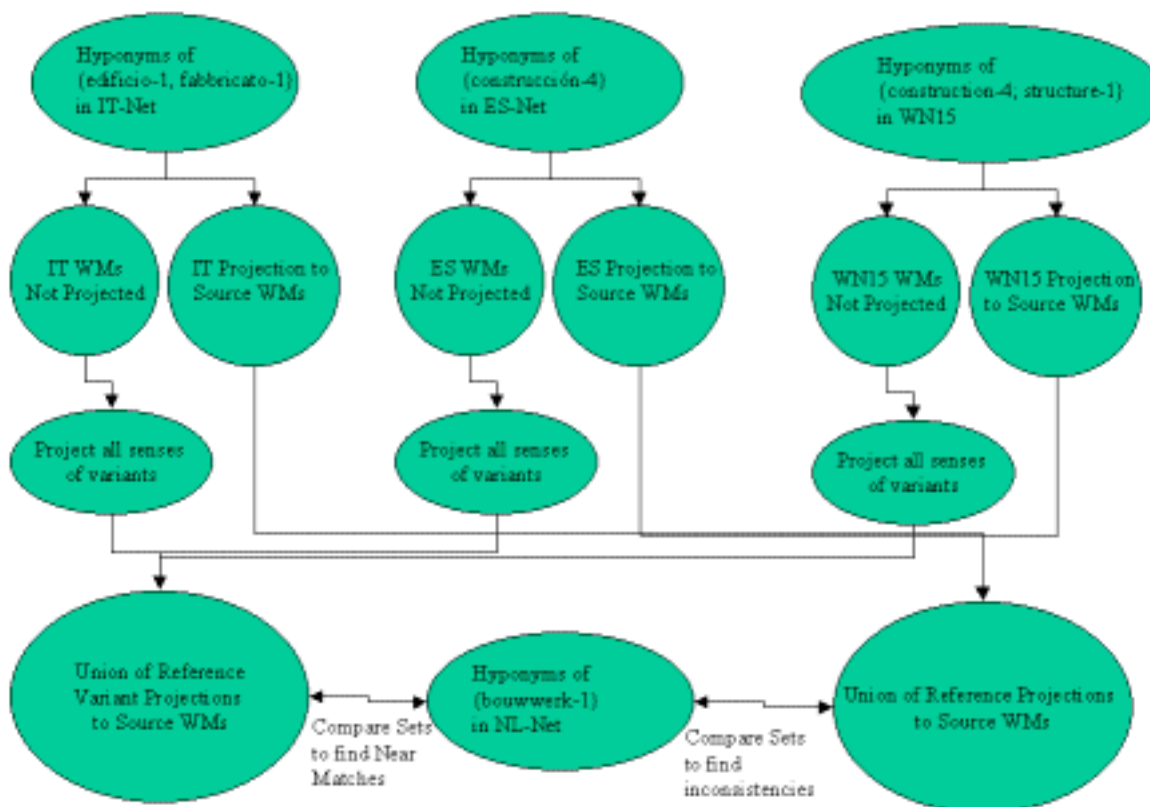


Figure 4: Comparison of Projected clusters of word meanings (WMs) to the Dutch wordnet.

In Appendix I, each site reports on the comparison of their source wordnet for the assigned clusters by comparing it to the other Reference wordnets. Here we will summarize the results.

The selected hyperonyms that have been used to derive a semantic field are given in the next table. Each hyperonym is represented by a synset, rounded by curled brackets, where we listed only a single variant. In some cases, several hyperonyms are given to represent the field. A field could not always be represented. For example, for Dutch there is no equivalent for *container*. After each hyperonym or set of hyperonyms we have listed the number of (sub)-hyponyms that occur in the wordnets. Finally, in the case of PLANT and ANIMAL the sets in WordNet1.5 are extremely big. We therefore limited PLANT to the first 3 hyponymy levels only and ANIMAL to the major classes such as MAMMAL, BIRD. The selections in the other wordnets are however complete. Because of the size of the field HUMAN we have split it into two sub-fields: ARTIST and WORKER.

Table 28: Hyperonyms in the wordnets selected for the in-depth comparison

	WN		NL		IT		ES	
BUILDING	{construction-4}	1220	{bouwwerk-1}	344	{costruzione-1} {edificio-1} {manifattura-1} {dimora-2}	194	{construcción-4}	548
COMESTIBLE	{food-1}	2123	{voedsel-1}	151	{cibo-1}	157	{alimento-1}	533
CONTAINER	{container-1}	567	{bak-1} {bergplaats-1}	26	{contenitore-1}	161	{contenedor-2}	245
COVERING	{covering-4} {covering-5}	1024	{bedekking-1}	27	{involucro-1} {copertura-2}	40	{cubierta-1} {cubierta-7}	425
GARMENT	{wear-1} footwear-1 {garment-1}	277	{kledingstuk-1}	23	{indumento-1}	156	{indumentaria-1} {calzado-1}	215
PLACE	{location 1}	1881	{plaats-1}	533	{luogo-1} {luogo-2}	54	{lugar-1}	373
FURNITURE	{furniture-1}	174	{meubelstuk-1}	11	{mobile-2}	75	{mobiliario-1}	65
PLANT	{plant-1} (first 3 levels)	802	{plant-1} {gewas-1}	28	{pianta-1}	474	{planta-1}	467
ANIMAL	{ animal-1 } (major levels)	2017	{dier-1}, {gedierte-1}	26	{animale-1}	563	{ animal-1}	682
ARTIST	{artist-1}	71	{kunstenaar-1} {artiest-1}	4	{artista-1}	91	{artista-2}	30
WORKER	{worker-2}	675	{werknemer-1}	9	{lavoratore-1}	552	{trabajador-1}	356
INSTRUMENT	{instrument 2}	509	{instrument-1}	437	{strumento-1}	867	{herramienta-1}	185
VEHICLE	{ vehicle 1}	410	{voertuig-1}	21	{veicolo-1}	172	{transporte-5}	189
FEELINGS	{feeling-1}n {experience-6}v {feel-7}v {feel-8}v	448	{voelen-4}v {voelen-5}v {gevoel-2}n {gevoel-3}n	87	{sentimento-1}n {percepire-1}v {provare-7}v	178	{sentimiento-1}n {sensación-6}n {sentir-3}v {sentir-5}v	253
PHENOMENA	{phenomenon-1}n	1012	{verschijnsel-1}n	353	{fenomeno-1}n	100	{fenómeno-1}n {caer-57}n	415
COOKING	{cook-1}v {cook-2}v {cook-3}v {cook-4}v {cooking-1}	57	{klaarmaken-2}v {koken-2}v	3	{cuocere-1}v, {preparare-3}v, {cuocere-2}v	24	{cocer-3}v {cocina-1}n	20
SOUNDS	{sound-13}v {sound-5}n {utter-3}v	271	{geluid-2}n {klinken-2}v	22	{rumore-1} {emettere-3}	45	{sonido-2} {sonar-3}v {emitir sonidos-1} v	139
MOVEMENT	{motion-1}n {motion-2}n {motion-5}n {move 1}v {move 4}v {move 2}v	1891	{beweging-1} {bewegen-1} {bewegen-2}	1313	{movimento-1}n {muoversi-1}n {muovere-1}v	148	{movimiento-8}n {movimiento-2}n {movimiento-1}n {mover-1}v {mover-3}v {moverse 4}v	600
KNOWLEDGE	??	??	{kennis-2}n {weten-2}v	53	{conoscenza-3}v {disciplina-1}n {conoscere-1}v	223	{información-1}n {pensamiento-2}n {teoría-3}v {disciplina-2}n {pensamiento-1}n	159

For each cluster or for major hyperonyms within a cluster, the following data is given:

1. the number of hyponyms in each wordnet linked to the given hyperonyms and give an overview of the equivalence relations per equivalence type.
2. the concepts that occur in both the Source and Reference wordnets but have different classifications across the wordnets.
3. the concepts that cannot be projected from the Reference wordnets to the Source wordnet.

In the next table the results are listed for each cluster. The first column gives the number of (sub)-hyponyms in each field (WM). The second column for each wordnet gives the number of ILI-records that is linked to the (sub)-hyponyms or descendants. Note that there can be multiple ILIs for a single synset, that some synsets have no ILI and that different synsets may share the same ILI. There is thus a many-to-many mapping between the ILIs and the synsets and there can be less or more ILIs than descendants. Note that we omitted the number of ILIs for WordNet1.5 because it is currently the same. The third column gives the number of WMs that intersected (\cap) with the Source wordnet WMs for each semantic field. The **source** set is marked in the table. The fourth table then represents the percentage of this intersection of the total number of WMs in the source wordnet. The last column in the table gives the number of near-matches (NM) that have been recovered for each field.

Table 29: Projections and Intersections of comparing the First Subset

	WordNet1.5			Dutch Wordnet				Spanish Wordnet				Italian Wordnet				N M
	WM	\cap	$\cap\%$	WM	ILIs	\cap	$\cap\%$	WM	ILIs	\cap	$\cap\%$	WM	ILIs	\cap	$\cap\%$	
building	1220	170	48%	351	223	X	X	553	548	133	38%	470	7	15	4%	34
comestible	2156	103	54%	192	154	X	X	541	533	70	36%	157	51	24	13%	19
container	567	15	58%	26	31	X	X	245	245	14	54%	161	7	8	31%	3
covering	1024	18	67%	27	43	X	X	431	425	17	63%	40	3	5	19%	10
garment	277	127	59%	23	36	23	11%	215	215	X	X	156	3	3	1%	14
place	1881	373	100%	533	425	58	16%	373	373	X	X	54	26	16	4%	54
furniture	174	65	100%	11	15	5	8%	65	65	X	X	75	4	4	6%	6
plant	802	132	28%	28	40	19	4%	467	468	X	X	474	261	170	36%	23
animal	2017	311	55%	26	43	156	28%	682	682	361	64%	563	318	X	X	84
artist	71	3	3%	4	4	11	12%	30		2	2%	91		X	X	65
worker	675	287	52%	9	146	146	26%	356		262	47%	552		X	X	51
instrument	509	215	24%	437	266	25	3%	185	185	29	3%	867	393	X	X	8
vehicle	410	106	62%	21	35	15	9%	189	189	94	55%	172	72	X	X	16
feelings	456	25	28%	89	139	X	X	256	253	17	19%	179	32	9	10%	29
phenomena	1020	16	4%	356	241	X	X	419	415	10	3%	100	23	6	2%	17
cooking	57	20	100	3	6	1	5%	20	20	X	X	24	13	7	35%	1
sounds	271	139	100	22	33	16	12%	139	139	X	X	45	41	19	14%	63
movement	1891	95	64%	1313	1304	90	61%	600	600	64	43%	148	98	X	X	89
knowledge ⁷	X	X	X	53	69	16	7%	159	159	17	8%	223	15	X	X	X

⁷ The comparison for the field knowledge could not be completed due to a technical problem.

The percentage of intersection directly reflect compatibility of wordnets. Extreme high intersection is found between WordNet1.5 and the Spanish wordnet, as can be expected given their methodology: PLACE, FURNITURE, COOKING, SOUND. However, many other projections still give a reasonable result around 50%, given the fact that we have not reached full coverage. If we look at intersections below 10% we first of all see that the Dutch wordnet lacks coverage (FURNITURE, PLANT, VEHICLE, COOKING, KNOWLEDGE) and that the Italian wordnet lacks equivalence relations (BUILDING, GARMENT, PLACE, FURNITURE). Both conclusions have already been suggested by previous data. A new fact is that PHENOMENA projected from WordNet1.5, ES and IT to Dutch, gives an extremely low intersection but still represents a large cluster in Dutch. This may point to a significant difference in classification in the Dutch wordnet. The following alternative classifications have been found for concepts which are PHENOMENA in the Reference wordnets (WordNet1.5, the Spanish wordnet and the Italian wordnet) but not in the Dutch wordnet:

- process/ change/ condition proces-2; verandering-1; gesteldheid-1 (all more general)
- systems: systeem (mechanisme)
- weather: weersgesteldheid (weather condition)
- power/force: energie-2 -> kracht-6 -> vermogen-; krachtveld
- possibilities: mogelijkheid
- diseases: ziekte-1

The alternative classifications show that there are many possibilities to describe a situation, which are not incompatible.

By further inspecting classification differences and the kind of equivalence mapping of all the fields we have come to the following conclusions:

1. Most mistakes are due to wrong translations, only a few are due to wrong classifications.
2. Alternative classifications occur quite regularly:
 - constructions: movable constructions; parts of buildings; institutions
 - comestibles: products such as fruits, grain, corn, seeds; drinks; parts
 - containers: object
 - covering: garments; parts of garments
 - feelings: stimulus (cause to feel like); more general experiences; attitudes; abilities
 - phenomena: process/ change/ condition; systems; weather conditions; power/force; possibilities; diseases
 - furniture: artifact or object
 - places: imaginary places; geographic terms; facility/installation (e.g. sports fields); containers
 - plants: microorganism; vegetables
 - sounds: communicate, breathe
 - cooking: creation, change
 - movement: sport; natural phenomena
3. There are quite a few cases of regular polysemy (e.g. constructions and installations or facilities) which can be resolved by conflating word meanings in the ILI.
4. ES only has eq_synonym relations, while NL and IT also have other types of equivalences.
5. ES correspondence to ILI is practically one-to-one, while NL tends to have more ILIs than synsets; and IT less ILIs than synsets.
6. ES tends to have more synsets than the others —at this stage of the project --- covering most of the synsets projected from other the WNs.

3.2 Overall comparison of Subset1

3.2.1 Introduction

The main objective of the overall comparison is to measure the degree of coverage and intersection of subset1. The statistics have been extracted at three levels:

- 1) Individual level (data provided by each site without any cross comparison).
- 2) Degree of coverage of WN1.5.
- 3) Overlapping with the other sites.

For this comparison each site (NL, IT, SP) has generated two sets (one for nouns and one for verbs) of hyponymy chains. For example, the next list of Dutch senses is generated for "opstijgen" (take off) by recursively taking all the hyperonyms:

- opstijgen (take off) stijgen (move to a higher position) verplaatsen (move location) voortbewegen (move location) bewegen (move reflexive) bewegen (move intransitive) veranderen (change)

To be able to compare these chains, each word sense in the chain has been replaced by eq_synonym and eq_near_synonym relations. When we reverse this chain (from top to bottom) we get the following result:

00064108-v 01046072-v 01046072-v 01046072-v 01055491-v 01094615-v 00257753-v

This means that the Dutch equivalent to ILI record number 00064108-v has as hyponym (the equivalent to ILI record number v 01046072-v) and this one has as hyponym (the equivalent to ILI record number 01046072-v), etc. Note that multiple translations lead to different chains.

In some cases (all of them for Dutch and Italian) an ILI chain contains nodes that have not been linked to WN1.5 equivalents. In these cases the original word instead of the ILI record number was used to identify the node. We then derived two statistics (when the differences are relevant) for chains with and without the untranslated nodes.

Two kinds of measurements have been developed: sense-based (synset or ILI) and chain based. Furthermore, the chain-based measurements have been divided into node-coverage and edge-coverage:

- Edge-coverage of chains means that not only the synsets but also the hyponymy relations between them are covered by the different wordnets.
- Node-coverage of chains means that the synsets are covered but not the hyponymy relations. Perhaps another relation holds between the corresponding synsets or perhaps they are unrelated.

Consider, for instance, that languages L1 and L2 contains the following ILI chains:

L1:

1--2--3
and
4--5

L2

1--2
3--4--5

The chain 1--2--3--4--5 is node-covered by both L1 and L2 languages but is not completely edge-covered by any of them. There are, however, two subchains of length 3, one for each language, and 2 subchains of length 2, one for each language too.

Both measurements are important and can be used in different way. Of course edge-coverage is more difficult to achieve (covering an edge implies covering the two related nodes and the relation between them -in the same direction-). A high degree of edge-covering overlap means that the overlapping concepts exist and are lexicalized in all the

languages that overlap and that their structural (hyponym/hyperonym) relationships hold in the same way for such languages. A lower level of edge-covering overlapping could indicate:

- a) incompleteness in covering the nodes (can be measured by node-coverage)
- b) incompleteness of relations in the language (can be measured by edge-coverage)
- c) A genuine difference between languages

What must be done is:

- 1) assuring a high level of overlap in node-coverage
- 2) check if a possible lower edge-coverage is due to either b) incompleteness (to be corrected) or c) real differences.

Complete overlapping of chains (either at edge or node level) is difficult to achieve in the case of huge differences in the size of the wordnets to be compared (e.g. the nouns in the Spanish wordnet, which is the largest subset, still hardly covers 30% of the nouns in WN1.5). We have therefore developed two other kind of measurements that are more useful for comparison: subsequences and subsequences with gaps:

- Subsequences are simply chains of nodes/edges that exactly match a fragment of another chain. Subsequences can be classified according to their length.
- Subsequences with n gaps are chains of nodes/edges that match a fragment of another chain but failing to match n nodes of edges.

For example:

- Node subsequence of length 2:
Sequence:
00002728 00004865 05839075 06193747
Subsequence:
00004865 05839075
- Edge subsequence of length 2:
Sequence:
00002728 00004865 05839075 06193747
Subsequence:
00004865 05839075 06193747
- Node subsequence of length 3 with 1 gap:
Sequence:
00002728 00004865 05839075 06193747
Subsequence:
00004865 06193747
- Edge subsequence of length 4 with 2 gap:
Sequence:
00002728 00004865 05839075 06193747 01137195
Subsequence:
00002728 00004865 06193747 01137195

Subsequences with 1 and 2 gaps are reported here. Although other cases can be computed in an easy way, their usefulness is clearly lower.

The procedure we have developed in order to extract the statistics consists of four steps:

1. One of the WNs is taken as base. The set of chains is read and a graph structure (in fact a DAG) is built.
2. The other WNs are projected over this skeleton. Possible cycles are not allowed. All the nodes are incorporated into the graph but only the compatible edges are added (i.e. the graph can be extended with additional nodes, some of the existing nodes can be marked as covered by the new language and some of the edges too, new edges can be added but only in the case they don't produce cycles).
3. The graph once completed is fully traversed in order to generate all the paths covering it (from tops to leaves). The set of paths is written into a file.
4. The file is queried in a variety of ways for extracting the statistics.

This procedure has been carried out 4 times, taking each wordnet as a starting point: WN1.5, Dutch-WN, Italian-WN and Spanish-WN. Next, we can query the database in a normal or verbose way (see also Appendix II). When using the verbose mode, not only the number but also the actual occurrences of the overlapping cases are extracted. Some of these instances are presented as examples for each kind of query. Normal mode is used here for presenting the results and extracting some conclusions from them in order to improve the overlapping of the different wordnets. Verbose mode will be used in a 2nd step for selecting ILI nodes or edges partially uncovered and guiding the extension of subset1.

Each site has generated an ASCII file containing the ILI-chains for all synsets in their Subset1. The following statistics can be extracted:

1. Frequencies and ratios of chains / length / language
2. Frequencies and ratios of ILI records / language
3. chains completely overlapping for pairs of languages (6), triples (3) and all 4 languages (if any). (occurrences, frequencies and ratios)
4. The same as 3) but for sub-chains of length N, for any value of N.
5. The same as 4) but allowing a maximum of M gaps (non overlapping elements within the subchain). These gaps could be contiguous or not.

In appendix II, we have listed the programs and software utilities that have been used to extract the data.

3.2.2 Evaluation of individual wordnets

The next results are taken from an analysis of individual chain sets. We performed the whole process for those ILI-records having WN1.5 equivalents. The figures are computed prior to any process. The content of the different columns is the following:

ILI nodes: number of different ILI nodes appearing in the sets (only ILI numbers having a WN1.5 translation have been taken into account because the comparison in the other cases is at this moment impossible)⁸

tops: nodes with no hypernym link

leaves: nodes with no hyponym link

internal nodes: nodes not being tops or leaves

edges: number of edges appearing in the sets

chains: number of chains appearing in the sets

Table 30: ILI chains for nouns

	<i>ILI nodes</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>CHAINS</i>
ES	18577	11	14208	4358	18885	16784
IT	1608	18	1446	161	2390	5083
NL	5098	14	5091	1510	6124	14673
WN15	60557	11	47110	13436	61123	53467

Table 31: ILI chains for verbs

	<i>ILI nodes</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>CHAINS</i>
ES	3218	368	2377	673	2863	2393
IT	541	39	447	60	790	849
NL	2142	13	2135	807	3351	5136
WN15	11363	573	8446	2580	10816	8486

⁸ in some cases an ILI node appears both as leaf and as internal node, so Tops, Leaves and Internal nodes are not disjoint sets. So the result of adding the Tops, Leaves and Internal nodes could be greater than the number of ILI nodes.

Table 32: *ILI chains for nouns and verbs*

	<i>ILI nodes</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>CHAINS</i>
ES	21795	379	16585	5031	21748	19177
IT	2149	57	1893	221	3180	5932
NL	7240	27	7226	2317	9475	19809
WN15	71920	584	55556	16016	71939	61953

The next tables show the same data as before but⁹:

- 1) Removing chains including inconsistent information, e.g. for nouns, those chains with verbs or adjectives.
- 2) Removing garbage (null lines, etc.).
- 3) Replacing repeated ILI-nodes in a chain for only one occurrence.
- 4) Removing those chains included in another ones (the hypernym chains).
- 5) Removing the repeated chains.

Table 33: *ILI chains for nouns*

	<i>ILI nodes</i>	<i>%WN</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>%WN</i>	<i>CHAINS</i>
ES	18577	30.68	11	14208	4358	18885	30.90	16784
IT	1608	-	17	1444	161	2390	-	4671
IT2(*)	1115	1.84	17	977	135	1664	2.72	3197
NL	5090	-	13	5083	1419	5883	-	14661
NL2(**)	3806	6.28	13	3807	1129	4491	7.35	14661
WN15	60557	100	11	47110	13436	61123	100	53467

(*) **IT2**: is Italian WordNet without exclusive Italian ILI Records.

(**) **NL2**: is Dutch WordNet without exclusive Dutch ILI Records.

Table 34: *clean ILI chains for verbs*

	<i>ILI nodes</i>	<i>%WN</i>	<i>Tops</i>	<i>Leaves</i>	<i>Internal Nodes</i>	<i>EDGES</i>	<i>%WN</i>	<i>CHAINS</i>
ES	3218	28.32	368	2377	673	2863	26.47	2393
IT	541	-	39	445	60	790	-	783
IT2(*)	437	3.85	39	359	41	634	5.86	629
NL	2085	-	12	2080	743	3088	-	4968
NL2 (**)	1781	15.67	12	1776	673	2690	24.87	4968
WN15	11363	100	573	8446	2580	10816	100	8486

(*) **IT2**: is Italian WordNet without exclusive Italian ILI Records.

(**) **NL2**: is Dutch WordNet without exclusive Dutch ILI Records.

Some conclusions can be extracted from these results:

- 1) There is a (relatively) high number of Italian and Dutch nodes without WN1.5 equivalent (31% of Italian nouns, 25% of Dutch nouns, 19% of Italian verbs, 14% of Dutch verbs). Without equivalents for these nodes, complete overlapping is not possible, not only at the node level but also, to a great extent, for the chains including such nodes.
- 2) There is (in general) a slightly better coverage in terms of edges than in term of nodes. This is a positive result because it seems to indicate that, when present, nodes are connected in a consistent way.

⁹For Italian and Dutch we present the data in two forms: raw data and chains where nodes without WN1.5 equivalent are dropped out.

The next two tables present the number and % of noun and verb chains classified by length for each language.

Table 35: Frequencies and ratios of noun chains / length / language

	WN		NL		IT		ES	
	frequency	%	frequency	%	frequency	%	frequency	%
1	0	0	5	0.04	0	0	0	0
2	33	0.06	21	0.17	642	13.63	86	0.51
3	521	0.97	509	4.14	1662	35.29	761	4.54
4	2220	4.15	1423	11.59	445	9.45	1678	10.02
5	5664	10.59	2306	18.78	925	19.64	3088	18.44
6	12730	23.81	2656	21.63	543	11.53	4010	23.95
7	11741	21.96	2708	22.05	304	6.46	3565	21.29
8	8737	16.34	1651	13.44	188	3.99	2204	13.16
9	5940	11.11	708	5.76	0	0	849	5.07
10	3305	6.18	208	1.69	0	0	344	2.05
11	1400	2.62	56	0.46	0	0	121	0.72
12	517	0.97	24	0.20	0	0	27	0.16
13	364	0.68	5	0.04	0	0	11	0.07
14	213	0.4	2	0.02	0	0	0	0
15	75	0.14	0	0	0	0	0	0
16	7	0.01	0	0	0	0	0	0
Total	53467	100	12282	100	4709	100	16744	100
Average	7.19		6.22		4.15		6.22	

Table 36: Frequencies and ratios of verb chains / length / language

	WN		NL		IT		ES	
	frequency	%	frequency	%	frequency	%	frequency	%
1	236	2.79	3	0.01	0	0	200	8.36
2	1837	27.72	102	0.43	451	42.42	765	31.97
3	2530	29.92	260	1.09	90	8.47	676	28.25
4	1959	23.17	838	3.52	125	11.76	436	18.22
5	1029	12.17	1328	5.58	201	18.91	218	9.11
6	462	5.46	1977	8.31	142	13.36	73	3.05
7	250	2.96	2284	9.60	42	3.95	23	0.96
8	109	1.29	2520	10.59	12	1.13	2	0.08
9	32	0.38	2562	10.77	0	0	0	0
10	10	0.12	2769	11.64	0	0	0	0
11	2	0.02	2512	10.56	0	0	0	0
12	0	0	2152	9.05	0	0	0	0
13	0	0	1571	6.60	0	0	0	0
14	0	0	1080	4.54	0	0	0	0
15	0	0	779	3.27	0	0	0	0
16	0	0	532	2.24	0	0	0	0
17	0	0	277	1.16	0	0	0	0
18	0	0	132	0.55	0	0	0	0
19	0	0	68	0.29	0	0	0	0
20	0	0	31	0.13	0	0	0	0
21	0	0	9	0.04	0	0	0	0
22	0	0	1	0.00	0	0	0	0
Total	8456	100	23787	100	1063	100	2393	100
Average	3.58		9.59		3.69		3.01	

Conclusions extracted from these tables:

- 1) the distribution in the case of nouns appears to be quite nice, at least it follows quite closely WN1.5 distribution. The differences in length (we can use the average length as measure) can be explained by the lower coverage (e.g. 7.19 in the case of WN1.5 vs 6.22 in the case of Spanish or Dutch).

- 2) In the case of Italian the average length is lower (4.15). The lower size of Italian subset1 can account partially for this difference. Another reason could be the concentration of Italian nominal nodes in the higher levels of the hierarchy or simply the lack of hyponymy/hypernym relations. The fact that with a nominal coverage significantly lower than the Spanish or Dutch (1608 ILI nodes vs 18577 or 5090) the number of tops is higher (17 vs 11 or 13) seems to point to this later explanation.
- 3) In the case of verbs the figures for Italian and Spanish wordnets are very close to WN1.5. This is not the case for the Dutch wordnet, which has long chains. A careful examination of some of these pathological chains shows that they come from wrong translation equivalences to WN1.5.¹⁰ If several (not many) of these equivalences appear in the higher levels of the hierarchy they have a multiplicative effect on the number of chains and the results appearing in the table are absolutely useless. The solution is to analyze carefully the most frequent long chains for checking the equivalence relations involved in them. There is no need to analyze all the long chains (i.e. those with length greater than 10) but to find the most frequent subsequences appearing in those long chains.

3.2.3 Global evaluation

The next three tables account for the coverage of the individual wordnets (NL, IT, SP), pairs (NL-IT, NL-SP, IT-SP) and full intersection (NL-IT-SP) against 1) WN1.5 and 2) the union NL-IT-SP

Table 37: Coverage of noun ILI records

	<i>Total</i>	<i>(60557)</i>	<i>(21828)</i>
	<i>frequency</i>	<i>% WN</i>	<i>% \cup(IT, NL, ES)</i>
ES	18577	30.68	85.11
IT	1608	2.66	7.37
NL	5090	8.41	23.32
\cap (ES, IT)	853	1.41	3.91
\cap (ES, NL)	2539	4.19	11.63
\cap (IT, NL)	389	0.64	1.78
\cap (ES, IT, NL)	334	0.55	1.53

Table 38: Coverage of verb ILI records

	<i>Total</i>	<i>(11363)</i>	<i>(4592)</i>
	<i>frequency</i>	<i>% WN</i>	<i>% \cup(IT, NL, ES)</i>
ES	3224	28.37	70.21
IT	541	4.76	11.78
NL	2085	18.35	45.41
\cap (ES, IT)	307	2.70	6.69
\cap (ES, NL)	918	8.08	19.99
\cap (IT, NL)	198	1.74	4.31
\cap (ES, IT, NL)	165	1.45	3.59

Table 39: Coverage of ILI records (total)

	<i>Total</i>	<i>(71920)</i>	<i>(26420)</i>
	<i>frequency</i>	<i>% WN</i>	<i>% \cup(IT, NL, ES)</i>
ES	21801	30.31	82.52
IT	2149	2.99	8.13
NL	7175	9.98	21.16
\cap (ES, IT)	1160	1.61	4.39
\cap (ES, NL)	3457	4.81	13.08
\cap (IT, NL)	587	0.82	2.22
\cap (ES, IT, NL)	499	0.69	1.89

It is difficult to extract conclusions from these three tables. An obvious objective is to increase the intersection between languages but it is difficult to achieve this objective in the same way we performed the construction of the Base Concepts. For one thing, we see that the union of all the ILIs (26420 nodes) almost represents the maximal set of

¹⁰ This is the result of the strategy to include the best two translations that have been generated by the automatic matching program. Since most verbs get several solutions a relatively high percentage of mistakes is generated.

synsets we aim at (30,000 synsets). Furthermore, not all these synsets are lexicalized in all the languages. However, we think that an improvement of the intersection can be obtained in an indirect way by filling gaps (or by extending subchains). These possibilities will be presented and discussed later.

The next ten tables account for the coverage of complete chains (at node and edge level) for nouns and verbs, projected over the different WNs. In the case of WN1.5, all the other WNs have been projected over while in the other cases WN1.5 has not been taken into account.

Table 40: Coverage of complete noun chains projected over WN1.5 structure

	<i>nodes</i> (53467)		<i>edges</i> (53467)	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
ES	7539	14.10	7539	14.10
IT	3	0.01	0	0
NL	288	0.54	4	0.01
\cap (ES , IT)	2	0.00	0	0.00
\cap (ES , NL)	150	0.28	3	0.01
\cap (IT , NL)	1	0.00	0	0.00
\cap (ES , IT , NL)	1	0.00	0	0.00

Table 41: Coverage of complete verb chains projected over WN1.5 structure

	<i>nodes</i> (8486)		<i>edges</i> (8486)	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
ES	1235	14.55	1235	14.55
IT	67	0.79	44	0.52
NL	413	4.87	86	1.01
\cap (ES , IT)	33	0.39	23	0.27
\cap (ES , NL)	133	1.57	28	0.33
\cap (IT , NL)	15	0.18	7	0.08
\cap (ES , IT , NL)	12	0.14	5	0.06

Table 42: Coverage of complete noun chains projected over Dutch wordnet

	<i>nodes</i> (12282)		<i>edges</i> (12282)	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
ES	3006	24.47	8	0.06
IT	0	0	0	0
\cap (ES , IT)	0	0	0	0

Table 43: Coverage of complete verb chains projected over Dutch wordnet

	<i>nodes</i> (23787)		<i>edges</i> (23787)	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
ES	1172	4.93	5	0.02
IT	21	0.09	0	0
\cap (ES , IT)	13	0.05	0	0

Table 44: Coverage of complete noun chains projected over Italian wordnet

	<i>nodes</i> (4709)		<i>edges</i> (4709)	
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
ES	2004	42.56	68	1.44
NL	238	5.05	45	0.96
\cap (ES , NL)	177	3.76	6	0.13

Table 45: Coverage of complete verb chains projected over Italian wordnet

	<i>nodes</i>	<i>(1063)</i>	<i>edges</i>	<i>(1063)</i>
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
ES	205	19.29	39	3.67
NL	135	12.70	17	1.60
\cap (ES, NL)	75	7.06	5	0.47

Table 46: Coverage of complete noun chains projected over Spanish wordnet

	<i>nodes</i>	<i>(16744)</i>	<i>edges</i>	<i>(16744)</i>
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
NL	380	2.27	8	0.05
IT	2	0.01	0	0
\cap (NL, IT)	1	0.01	0	0

Table 47: Coverage of complete verb chains projected over Spanish wordnet

	<i>nodes</i>	<i>(2397)</i>	<i>edges</i>	<i>(2397)</i>
	<i>frequency</i>	<i>%</i>	<i>frequency</i>	<i>%</i>
NL	297	12.39	89	3.71
IT	72	3.00	50	2.09
\cap (NL, IT)	25	1.04	14	0.58

The figures presented in the preceding tables are of rather limited use, since full coverage of the chains is not possible. The coverage in terms of complete chains is extremely low and the reason is the great differences in size between the different wordnets. Consider, for instance, the overlapping between WN1.5 and the Spanish wordnet. The ratio, when comparing nodes is for nouns 30.68%. When comparing full chains this figure drops to 14.10%. This is not necessarily a bad result. A possible interpretation could be that most of the coverage is concentrated in the highest levels of the hierarchy. This is confirmed by other evidence. Obviously better, and more useful, results will be obtained when dealing with incomplete sequences (both subsequences and sequences containing gaps). These cases will be considered in next.

The following four tables account for the overlapping of partial chains (node vs edge, noun vs verb) projected over WN1.5 structure, for different lengths of the chain

Table 48: Coverage of partial noun chains of NODES projected over WN1.5 structure

<i>LENGTH</i>	<i>ES</i>	<i>IT</i>	<i>NL</i>	\cap (ES, NL)	\cap (ES, IT)	\cap (IT, NL)	\cap (ES, IT, NL)	<i>WN</i>
1	53467	36102	52504	52376	36080	30912	30909	53467
2	47792	23597	39616	38828	23531	16181	16151	53467
3	45744	15219	24004	23448	15189	6767	6756	53434
4	41747	7896	14200	14009	7844	2048	2001	52913
5	23930	3709	6146	5659	3627	809	780	50693
6	23350	948	2888	2535	751	402	393	45029
7	11774	333	1095	901	265	236	228	32299
8	5007	93	435	374	63	11	9	20558
9	1795	14	73	67	9	0	0	11821
10	664	0	1	1	0	0	0	5881
11	137	0	0	0	0	0	0	2576
12	36	0	0	0	0	0	0	1176
13	7	0	0	0	0	0	0	659

Table 49: Coverage of partial noun chains of EDGES projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
1	47792	15185	29409	28928	15154	9168	9133	53467
2	45744	816	13647	13540	636	271	253	53434
3	41747	105	367	214	93	46	46	52913
4	23930	4	53	18	2	0	0	50693
5	23350	0	4	0	0	0	0	45029
6	11774	0	0	0	0	0	0	32299
7	5007	0	0	0	0	0	0	20558
8	1795	0	0	0	0	0	0	11821
9	664	0	0	0	0	0	0	5881
10	137	0	0	0	0	0	0	2576
11	36	0	0	0	0	0	0	1176
12	7	0	0	0	0	0	0	659

Table 50: Coverage of partial VERB chains of NODES projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
1	7861	6299	7188	6720	6248	5423	5382	8486
2	5273	1601	3334	2563	1483	840	765	8250
3	2980	233	1244	799	208	46	43	6383
4	1184	37	234	112	37	0	0	3853
5	107	0	24	8	0	0	0	1894
6	82	0	2	0	0	0	0	865
7	14	0	0	0	0	0	0	403
8	2	0	0	0	0	0	0	153

Table 51: Coverage of partial VERB chains of EDGES projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
1	5273	701	1406	1056	621	193	150	8250
2	2980	41	83	67	39	1	0	6383
3	1184	0	1	1	0	0	0	3853
4	107	0	0	0	0	0	0	1894
5	82	0	0	0	0	0	0	865
6	14	0	0	0	0	0	0	403
7	2	0	0	0	0	0	0	153

A short explanation is needed for interpreting these tables. In the case of node coverage a subsequence of length 1 corresponds to a simple node. As for WN1.5 all the sequences start with a top node, this means that for nouns every possible chain contains one of the 11 tops. In this way there are 53467 possible subchains of length 1 (i.e. one for each of the 53467 edges present in WN1.5). The fact that for Spanish the number of partial chains of length 1 is 53467 too, simply means that Spanish wordnet covers all of these 11 tops. In the case of Dutch the figure is a little lower and this means that at least one WN1.5 top is not covered by Dutch wordnet. The corresponding figures for verbs are lower simply because the number of WN1.5 tops is greater (573) and the degree of coverage for the different wordnets is obviously lower.

So, we must find a compromise between the degree of coverage and the significance of such coverage. The first rows present a high coverage but the significance is very low. As the length grows, the significance of the overlapping is greater but the coverage is poor. For nouns the interesting and useful figures can be found in the rows corresponding to lengths 3 to 7. In the case of verbs useful information can be obtained from lengths 2 and 3.

When considering the edge coverage we must taken into account that covering an edge means covering the adjacent nodes and the relation between them. So interesting information could be extracted from rows corresponding to edge lengths 2 or 3, for nouns, and 1 for verbs.

How to use this information? We can, by means of the verbose mode, select all the subchains of length 3 to 7 covered by 2 languages and try to complete these chains for the other language. For instance, for nouns, we can select the 901 chains of length 7 covered over WN1.5 by Spanish and Dutch wordnets. From these, 228 are covered too by Italian, so

only 673 chains must be checked for Italian. Fortunately most of these chains own a common prefix (i.e. some of the initial nodes of the chains are common to several of them). So the amount of work to be done is limited. In this way the information can be used for guiding the construction of subset2 trying to improve the overlapping between chains. Similar considerations could be pointed out for edge chains. In Appendix III, the tables are given for when the wordnets are projected over Dutch, Spanish and Italian. The results and conclusions are similar to the above.

Considering subsequences, as has been pointed out above, is a very useful source of information for 1) assessing the quality of coverage of the wordnets and 2) developing criteria for guiding the extension of the wordnets and obtaining data for supporting such criteria. Another complementary source of information consists of the subsequences containing gaps. We only discuss the results of overlapping subsequences with one or two gaps.¹¹ For one gap, all the tables are presented, for two gaps only the two tables corresponding to partial noun coverage projected over the WN1.5 structure are presented. The reason is simply that the conclusions are basically the same when projecting over the other wordnets and that for verbs the figures when dealing with rather long chains have little significance.

The following tables account for the overlapping of partial chains containing one gap (node vs edge, noun vs verb) projected over WN1.5, Dutch, Italian and Spanish WN structure, for different lengths of the chain.

Table 52: Coverage of partial noun chains of NODES with 1 gap projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
3	9553	5750	16836	15953	5460	5709	5672	53434
4	9293	4647	14210	13107	4314	4193	4127	52913
5	8742	3281	9753	9187	2988	2932	2901	50693
6	7721	1541	6373	5924	1284	270	227	45029
7	5831	589	2333	2041	496	12	11	32299
8	3853	147	664	499	103	3	3	20558
9	1682	35	190	78	19	3	3	11821
10	654	12	60	18	3	0	0	5881
11	164	1	16	10	0	0	0	2576
12	61	0	0	0	0	0	0	1176
13	7	0	0	0	0	0	0	659
14	2	0	0	0	0	0	0	446

Table 53: Coverage of partial noun chains of EDGES with 1 gap projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
3	0	1260	2969	2903	1192	334	309	52913
4	0	281	1478	1369	270	133	123	50693
5	0	50	273	184	47	45	45	45029
6	0	4	33	15	2	0	0	32299
7	0	0	5	0	0	0	0	20558

Table 54: Coverage of partial VERB chains of NODES with 1 gap projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
3	826	954	822	642	899	347	317	6383
4	486	111	317	220	104	2	2	3853
5	329	20	82	50	18	0	0	1894
6	121	0	8	2	0	0	0	865
7	39	0	1	0	0	0	0	403
8	6	0	0	0	0	0	0	153

¹¹ It is also possible to consider gaps of 3 or more nodes and even 0-gaps subsequences. The 0-gaps subsequences have gaps at the beginning and/or at the end of the chain. The usefulness of these sequences is however lower.

Table 55: Coverage of partial VERB chains of EDGES with 1 gap projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
3	0	0	11	6	0	0	0	3853

Table 56: Coverage of partial noun chains of NODES with 1 gap projected over Dutch wordnet

LENGTH	ES	IT	$\cap(ES, IT)$	NL
3	3694	3457	3428	12256
4	3543	1279	1243	11747
5	2987	471	456	10324
6	2222	222	221	8018
7	1333	179	177	5362
8	569	68	62	2654
9	151	2	0	1003
10	58	0	0	295
11	22	0	0	87
12	8	0	0	31

Table 57: Coverage of partial noun chains of EDGES with 1 gap projected over Dutch wordnet

LENGTH	ES	IT	$\cap(ES, IT)$	NL
3	1139	7	6	11747
4	516	0	0	10324
5	314	0	0	8018
6	73	0	0	5362
7	12	0	0	2654

Table 58: Coverage of partial verb chains of NODES with 1 gap projected over Dutch wordnet

LENGTH	ES	IT	$\cap(ES, IT)$	NL
3	13769	3385	2708	23682
4	13559	1671	1284	23422
5	12715	535	387	22584
6	10638	24	17	21256
7	8577	0	0	19279
8	6688	0	0	16995
9	4245	0	0	14475
10	2021	0	0	11913
11	1011	0	0	9144
12	520	0	0	6632
13	3	0	0	4480

Table 59: Coverage of partial verb chains of EDGES with 1 gap projected over Dutch wordnet

LENGTH	ES	IT	$\cap(ES, IT)$	NL
3	12	0	0	23422

Table 60: Coverage of partial noun chains of NODES with 1 gap projected over Italian wordnet

LENGTH	ES	NL	$\cap(ES, NL)$	IT
3	148	560	513	4067
4	139	239	226	2405
5	104	42	35	1960
6	15	12	10	1035
7	0	4	4	492

Table 61: Coverage of partial noun chains of EDGES with 1 gap projected over Italian wordnet

LENGTH	ES	NL	$\cap(ES, NL)$	IT
3	43	68	2	2405
4	1	0	0	1960

Table 62: Coverage of partial VERB chains of NODES with 1 gap projected over Italian wordnet

LENGTH	ES	NL	$\cap (ES, NL)$	IT
3	443	211	209	612
4	379	190	131	522
5	213	81	38	397
6	57	10	3	196
7	4	0	0	54

Table 63: Coverage of partial VERB chains of EDGES with 1 gap projected over Italian wordnet

LENGTH	ES	NL	$\cap (ES, NL)$	IT
3	0	2	2	522

Table 64: Coverage of partial noun chains of NODES with 1 gap projected over Spanish wordnet

LENGTH	NL	IT	$\cap (NL, IT)$	ES
3	5410	1931	1859	16658
4	4508	1577	1381	15897
5	3028	1026	986	14219
6	1698	361	67	11131
7	699	180	8	7121
8	186	42	1	3556
9	49	9	1	1352
10	9	1	0	503
11	1	0	0	159

Table 65: Coverage of partial noun chains of EDGES with 1 gap projected over Spanish wordnet

LENGTH	NL	IT	$\cap (NL, IT)$	ES
3	997	341	122	15897
4	466	91	34	14219
5	76	14	9	11131
6	10	5	0	7121

Table 66: Coverage of partial VERB chains of NODES with 1 gap projected over Spanish wordnet

LENGTH	NL	IT	$\cap (NL, IT)$	ES
3	205	263	101	1432
4	87	41	2	756
5	21	4	0	320
6	2	0	0	100

Table 67: Coverage of partial VERB chains of EDGES with 1 gap projected over Spanish wordnet

LENGTH	NL	IT	$\cap (NL, IT)$	ES
3	5	0	0	756

The following tables account for the overlapping of partial noun chains containing two gaps (node vs edge) projected over WN1.5 Tables with more gaps are not included here but can be easily generated, in the way described in 3.2.2.

Table 68: Coverage of partial noun chains of NODES with 2 gaps projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
4	7206	1654	6337	5620	1546	782	662	52913
5	2713	1548	6729	5939	1423	899	793	50693
6	2521	916	5502	4731	806	478	396	45029
7	2172	495	2176	1676	395	129	119	32299
8	1662	323	1249	1009	273	36	28	20558
9	1239	136	191	157	100	0	0	11821
10	606	30	50	50	26	0	0	5881
11	271	2	9	3	0	0	0	2576
12	179	1	1	0	0	0	0	1176
13	115	0	0	0	0	0	0	659
14	25	0	0	0	0	0	0	446
15	2	0	0	0	0	0	0	82

Table 69: Coverage of partial noun chains of EDGES with 2 gaps projected over WN1.5 structure

LENGTH	ES	IT	NL	$\cap(ES, NL)$	$\cap(ES, IT)$	$\cap(IT, NL)$	$\cap(ES, IT, NL)$	WN
4	2332	633	821	695	626	0	0	50693
5	2189	59	840	713	13	0	0	45029
6	1859	0	312	263	0	0	0	32299
7	1381	0	118	115	0	0	0	20558
8	902	0	2	2	0	0	0	11821
9	287	0	0	0	0	0	0	5881
10	86	0	0	0	0	0	0	2576
11	20	0	0	0	0	0	0	1176
12	2	0	0	0	0	0	0	659

The same considerations pointed out when presenting the partial chains are valid here. Obviously the chains are longer in this case (e.g. for having a subchain with one node gap the minimum length is 3, with 2 gaps the minimum length is 4).

A possible useful criterium for using this information could be simply generating the set of sequences belonging to the intersection of all languages and having a gap and then from this set generating for each language L the subset of sequences where this gap is not covered for L . For instance, for nouns, the row corresponding to length 5 chains shows that the intersection consists of 2901 sequences with 1 gap. In the worst case all the gaps correspond to every language, i.e. the gaps are not covered by any of the 3 languages other than English, and all the gaps correspond to different ILI nodes. In this worst case the task would consist of filling these 2901 gaps for each language. Doing so, the number of length 5 noun chains without gaps will increase from 780 to $780 + 2901 = 3681$ with an increment of 470%! If we consider also the length-5 noun chains with 2 gaps, another increment (in this case of 793 chains) can be obtained.

A careful evaluation of the most promising chains in order to optimize the human resources must be done, but we are sure that using the information provided by these tables and generating the corresponding occurrences (with the verbose option) can be helpful as a guideline for extending and improving the wordnets.

4. Updating the ILI

The ILI will be updated in 3 ways (see [Peters et al, fc] for more details):

1. Improving the information given for the current ILI-records: e.g. adding missing glosses
2. Adding missing concepts or gaps in the ILI occurring in the other wordnets
3. Adding globalized sense, grouping closely related senses of words in the ILI

Gaps (2) will be added later in the project. For Subset1 we have focussed on improving the glosses (1) and creating globalized sense-groups (3). The latter are needed to deal with the reduction of the high level of granularity of the WordNet sense distinctions and the inconsistent treatment of regular polysemy in lexical resources. As explained in [Peters et al, 1998], differences in the sense-distinctions across resources may lead to mismatches or fuzzy-matching across wordnets. For example, *university* may be used to refer to both the *institute* and the *building* but we see that resources often only represent one of these meanings, or conflate them in a single meaning. This may again result in a situation that the local synsets for *university* cannot be matched across wordnets.

To limit this danger, we extend the ILI with globalized senses that represent sets of more specific but related senses of the same word. Three main types of relations are distinguished:

1. metonymy, e.g. grouping building-institute senses
2. generalization, e.g. grouping specific uses of a single more general sense
3. diathesis alternation, e.g. grouping causative/inchoative senses

In Figure 5, we see that the original linking of Dutch, Italian and Spanish equivalents for *university* has been extended with an HAS_EQ_METONYM relation to a new globalized ILI-record *university* which contains a reference to two more specific meanings. Via the HAS_EQ_METONYM relations the synsets can be retrieved despite of the different ways in which they are linked to the more specific synsets. It is not necessary that the metonymy-relation also holds in the local language. In this example only the Dutch wordnet has two senses that parallel the metonymy-relation in the ILI.¹² The Italian and Spanish example only list one sense (which may be correct or an omission in their resources). In the case of Spanish there are multiple equivalences to both senses of *university*, whereas the Italian synset is only linked to the *building* sense. The Spanish example is in fact equivalent to the new globalized ILI-record.

Similar globalized records are added for generalizations or verbal alternations such as causative and non-causative meanings: *he opens the door*, versus *the door opens* [Levin 1993]. In that case an HAS_EQ_GENERALIZATION or HAS_EQ_DIATHESIS relation will hold for synsets linked to more specific ILI-records that can be grouped in these ways. The generation of these equivalence relations is done fully automatically. After extending the ILI with more global concepts, the HAS-EQ_METONYM, HAS_EQ_GENERALIZATION or HAS_EQ_DIATHESIS will be automatically generated for all synsets which have at least one of the specific ILI-records in the globalized ILI-records as the target of an EQ_SYNONYM or EQ_NEAR_SYNONYM relation. There is no need for the local wordnet builders to consider each of these equivalence-extensions manually.

¹² The relation between these two Dutch senses is now also expressed via the metonymy-equivalence relation to the more global ILI-record. The globalized ILI-record may also create metonymic relations between different forms which represent the same semantic relation, such as *universiteit* (university institute) and *universiteitsgebouw* (university building) in Dutch.

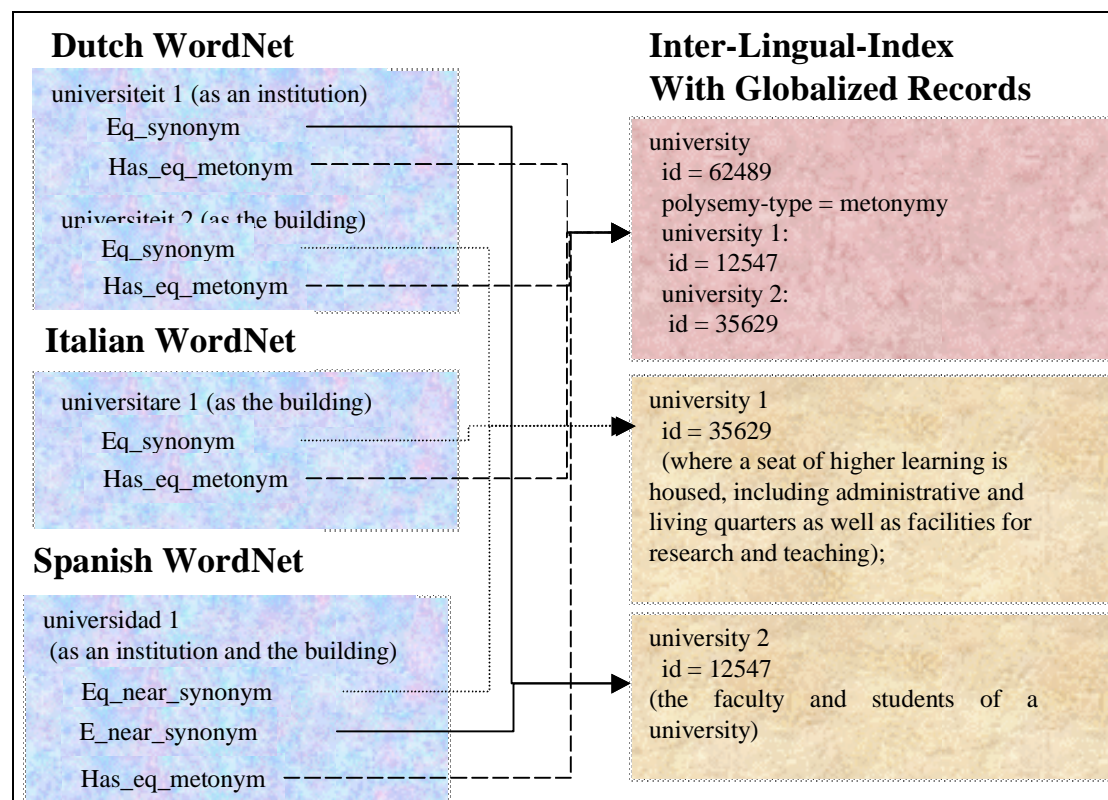


Figure 5: Inter-Lingual-Index with a globalized synset for *university*.

4.1 Clustering Methods

This section describes the clustering methods which have been and still are being applied in order to identify these sense groups.

4.1.1 Manual clustering

We started off with a manual examination of the polysemous words in the Base concept (BC) set and their senses which had originally been rejected in the BC selection process, but had been selected by at least two individual language partners. This set will be referred to below as the Rejected Concepts (RC). These originally rejected synsets were evaluated against the base concept set, and their possible inclusion into the BC set was investigated. Three different strategies have been applied in order to select relevant members from the RC set as new BCs.

1) The first strategy uses the average number of semantic relations selected noun base concepts have in WordNet1.5 (19.49) as the selection threshold for rejects. The following selection criteria have been applied:

- Because of the high number of direct hyponymic relations within the BC (636) direct RC hyponyms of existing BCs have principally not been selected. For instance, 'airplane' has not been selected because of its NBC direct hypernym 'aircraft'. 'rate' (a magnitude relative to a time unit; "they traveled at a rate of 55 miles per hour" or "the rate of change was faster than expected") is subsumed by BC 'relation' (an abstraction belonging to or characteristic of two entities or parts together) and has not been selected.
- Basic level concepts (43) like bed, wheel, shoe, window, glass, eye, soup, pants, antelope represent a level of lexicalisation which is too specific, and have not been selected.
- Taxonomic terms within the field of biology have not been selected as new BCs. They have very specific technical meanings, and are subsumed by the BC 'group'.

This first selection task yielded 14 new potential BC members.

2) The second selection method consisted of the manual examination of rejected synsets which share a word form member with one or more NBC synsets. Only word forms shared by at least four synsets have been taken into account and it has been investigated whether the rejected synsets belonging to these sense groups did not all originate from only one language specific wordnet, but have a more or less even distribution over the different language sites. This method resulted in 100 new BC members.

3) The third strategy used the metric developed by [Agirre & Rigau, 1996] for computing conceptual distance between WordNet nodes representing different senses of the same word. Using a threshold of .3 yielded 21 RC synsets as candidates for BC membership.

For the manual identification of encoding of metonymic regularities between senses the following aspects of systematic polysemy apply amongst others [Apresjan, 1974] , [Pustejovsky 1995]:

- a general notion of involvedness: the senses are related within a typical situation; e.g. social group versus belief, organisation vs building; result, e.g activity vs product;
- constituent or portion/part vs, whole relations, e.g wood-tree, person-social group
- function e.g. liquid-beverage

For the identification of generalization it was difficult to find clear cut criteria. The general criteria used included the level of fine-grainedness of the sense distinctions and the possibility to make an ontological generalization over the senses involved, which constitutes a lowest common denominator that all grouped senses share.

This first manual clustering round resulted in 31 verb and 119 noun groups.¹³

4.1.2 (Semi-)automatic clustering

In addition to the manual clustering, various automatic clustering methods have been examined. Most of these methods rely on the internal hierarchical organization of WordNet and, except for autohyponymy (see section 4.1.2.2), they are all used in the WordNet interface to compute semantic similarity. With respect to using external resources to aid clustering, we have only looked at CoreLex (section 4.1.2.5). However, we envisage using other existing lexical resources, such as machine-readable dictionaries and ontological classifications. Thus far, we have limited ourselves to homographs of the same part of speech. The methods are described in the following paragraphs.

4.1.2.1 Sisters

Word senses that share the same hypernym are called sisters¹⁴. In the example below, both senses of *table* have *furniture* as their direct hypernym:

table-2

'a piece of furniture having a smooth flat top supported by one or more vertical legs; "it was a sturdy table"'

table-3

'a piece of furniture with tableware for a meal laid out on it; "I reserved a table at my favorite restaurant"'

Using the sister criterion generates patterns of generalization. In the example above, the given senses can be used to refer to the same object, highlighting different aspects of it. However, in some cases the clustered senses refer to different objects in the real world. This is illustrated by the following example, where all three senses share the direct hypernym *vine*.

butterfly pea

- 'vine of tropical Asia having pinnate leaves and bright blue yellow-centered flowers'
- 'large-flowered wild twining vine of SE and C US having pale blue flowers'
- 'large-flowered weakly twining or prostrate vine of NJ to tropical E N America, sometimes cultivated for its purple and white flowers'

As these senses denote different species, they are not near-synonyms. However, they are very similar in nature, and can be clustered on that basis. It must be taken into account that, and this is true for all generalizations, the meanings cannot be used interchangeably. The most specific semantic content these particular senses share is the meaning of the direct hypernym.

¹³ SHE has also provided manually identified metonymy and generalization relations for the semantic fields Building and Institute, which resulted in 33 new groups. This was done to investigate effectiveness of metonymic clusters.

¹⁴ The sister relation is not limited to two senses, but can also occur between three or more senses of the same word. Sometimes, a particular word exhibits more than one type of sister relation.

4.1.2.2 Autohyponymy

The term *autohyponymy* is used to refer to words whose senses are each others direct hypernyms or hyponyms (Cruse, 1986). Sharing the same hypernymic chain (except for the first node) provides us with a number of combinations where the meanings are very similar and clustering results in homogenous groups. Look at the following examples, where the first sense is the most specific one:

- *variety-3, species*
'a specific kind of something: "a species of molecule" or "a species of villainy"'
- *variety-6, kind, sort, form*
'a category of things distinguished by some common characteristic or quality;
"sculpture is a form of art"; "what kinds of desserts are there?"'
- *understand-3, read, interpret, translate*
'make sense of a language; "She understands French"; "Can you read Greek?"'
- *understand-1*
'know and comprehend the nature or meaning of; "She did not understand her husband";
"I understand what she means"'

As this method also leads to generalization clusters it is the meaning of the hypernym synset that can be used to characterize the resulting sense cluster. The specific sense is subsumed by the general one; the hyponym carries extra meaning which is not shared by its parent and/or is typically used in a specific domain.

4.1.2.3 Twins

Twins are synsets that have at least three members in common as the example below illustrates. Their meanings are defined by 'of rules or patterns' and 'act in disregard of laws and rules', respectively.

- *violate, fail to agree with, go against, break-13, be in violation of*
- *violate, go against, breach, break-6, be in violation of*

This example seems to validate clustering on the basis of the twin criterion. However, some of the twin groupings are more problematic. The synsets below have the following incompatible glosses: 'motion that does not entail a change of location; "the reflex movements of his eyebrows revealed his surprise"' and 'the act of changing your location from one place to another'.

- *change of position, motion, movement, move-3*
- *change of location, motion, movement, move-4*

A number of synsets are linked by a twin relation only because they contain spelling variants, such as *sestet*, *sextet*, *sextette*. As we have not yet examined the twin relation in great detail, we cannot fully assess the validity of this method. However, it seems that even in cases where synsets only share two members, this can also be an indication that clustering is possible. An example is *travel-4, journey* and *travel-2, journey*, where the meanings are very closely related.

4.1.2.4 Cousins

WordNet1.5 contains a list of 105 node top pairs whose hyponyms exhibit a specific relation to each other (see WordNet database documentation on groups, file groups.⁷¹⁵). These pairs have been identified and listed by lexicographers. The treatment of these so-called cousins is still in its experimental stage; the resulting list is incomplete and does not offer a consistent and structured list of recurrent patterns between sets of words. Examples of cousin relations are *container-containerful* and *food-tableware*, listed below.

container-1

'something that holds things, especially for transport or storage'

containerful-1

'the quantity that a container will hold'

¹⁵ This documentation is included in the WordNet database, downloadable from <http://www.princeton.edu/~wn/>

food-1, nutrient

'any substance that can be metabolized by an organism to give energy and build tissue'

tableware-1

'articles for use at the table'

Looking at the first pair, there are a large number of words that occur both as hyponyms of the *container* node and the *containerful* node, such as *bag*, *can*, *cup*, *glass*, *shovel*, *spoon* and *thimble*. These are all good examples of the regular polysemic pattern that exists between *container* and *containerful*. On further investigation, we find that the cousin relation is not limited to senses sharing a word form. For example, WordNet contains no words that have both a *food* and a *tableware* meaning. While words such as *silver plate*, *gold plate*, *crockery* and *chop sticks* all occur as hyponyms of *tableware*, they are not found in a *food* sense. Cousin relations, thus, do not necessarily generate regular polysemous patterns, but sometimes capture semantic relations between words of a more schema-like nature. Within the scope of the present research we are only interested in sense distinctions of individual words and can only use those cousin relations generating clusters that share word forms.

4.1.2.5 CoreLex

An attempt at making systematic polysemic patterns in WordNet explicit has been made by Buitelaar (1998). The CoreLex database¹⁶ contains 126 semantic types, covering 39,937 nouns in 317 systematic polysemous classes. Three steps were taken to derive CoreLex from WordNet. Firstly, all polysemous nouns in WordNet were reduced to a set of Basic Types, corresponding largely to WordNet's 'unique beginners' and 'top nodes', such as *artifact*, *causal agent*, *shape* and *act*. Subsequently, systematic groupings of nouns were created on the basis of their Basic Types distributions. For example, the noun *banana*, occurring both in a *food* and a *plant* sense, was put in a group with other nouns exhibiting the same pattern, such as *coriander*, *grapefruit*, *plantain* and *mulberry*. The final step consists of integration into the Core Lexical Engine [Pustejovsky, 1995].

On examining the polysemous classes, we found a number of disadvantages to the CoreLex system. Firstly, 19 of them consist of only one Basic Type and therefore do not display systematic polysemy. More importantly, the generated classes are not always homogeneous in nature; particularly the larger groups do not necessarily exhibit regular polysemic patterns and occurrences of 'monsters' are not infrequent. Often there is scope for further subclustering. For example, we find *bundle*, *package*, *packet*, *ragbag*, *deck*, *edition*, *library*, *menagerie*, *repertory* belonging to the same CoreLex type (*arg*, a combination of the Basic Types *artifact* and *group*) where we find the first 4 words covered by the more specific hypernymic nodes *collection-1* and *container-1* and the last three by *collection-1* and *facility-1*. For our purposes, the main problems with CoreLex are caused by the fact that the Basic Types are largely based on very high-level nodes in the WordNet hierarchy. In order to obtain more homogeneous classes, we propose to examine recurrent distributional patterns at a more specific level in the hierarchy.

4.2 Testing automatically created sense groups

We have performed a first validation test of automatic procedures for deriving sense clusters. For this purpose we carried out an experiment in which different fragments of the Dutch and Spanish wordnets were compared, both before and after extending the ILI with composite ILI records. For the experiment, composite ILIs have been generated automatically on the basis of two methods:

- We selected a number of metonymic relations and subsequently extracted all words that have one sense occurring as a (sub)hyponym of one element of the relation and another sense as a (sub)hyponym of the other element. Some of these relations feature in the cousin table, discussed in section 4.2.4. As suggested in section 4.2.5, the selected relations generally consist of hypernymic nodes that are more specific than WordNet's top nodes and unique beginners. This method generates regular polysemic patterns.
- From the words selected by the above-mentioned method, we clustered those word senses that are (sub)hyponyms of one of the members of the metonymic relations selected in this experiment. This method extracts generalization clusters and extends the sister relation discussed in 4.2.1 to include those senses that are not direct hyponyms of the shared hypernymic node, i.e. senses that are not co-hyponyms. This method also subsumes autohyponymy.

¹⁶ Available from <http://www.cs.brandeis.edu/~paulb/CoreLex/overview.html>

Table 70 gives the totals for the extracted records. In total 700 new composite ILI-records have been added (214 metonymic groupings and 486 generalization pairs), involving 1557 ILI-records. Note that this method is fully automatic and can easily be extended to all senses in WordNet1.5. In general, we see here that the largest metonymic classes are *animal/food* and *plant/food*. The largest set of generalization is extracted for *move*. After extending the ILI with the new concepts, the equivalence relations of the Spanish and Dutch wordnet to the ILI have been updated. This is done automatically by the database: any synset that is related to an ILI-record included in a composite ILI will get an additional metonymy or generalization link to this composite ILI-record. For the Dutch wordnet 602 links have been added, and for the Spanish wordnet 521 links.

Table 70: Automatic derived generalizations and metonymy-relations

<i>Semantic Class</i>	<i>Total no Descendants</i>	<i>Generalization Clusters</i>	<i>Metonymic Clusters</i>	<i>Intersecting Metonymy Class</i>	<i>Total Composites</i>	<i>Percentual Coverage of all Senses</i>
animal ¹⁷	3842	80	81	food	161	4.19%
plant	4750	48	100	food	148	3.11%
food	2123	64	181	animal/plant	245	11.54%
organization	846	31	25	construction	56	6.61%
construction	1210	81	25	organization	106	8.76%
move	708	176	8	sound	184	25.98%
sound	192	6	8	move	14	7.29%
Total	13671	486	214		914	6.68%

To measure the effect, we mapped Spanish (ES) and Dutch (NL) fragments before and after extending the ILI with these records. All descendants of Dutch and Spanish representatives of the above classes were selected, e.g. all (sub)hyponyms of *bouwwerk-1* (*construction*) in Dutch and *construcción-4* (*construction*) in Spanish. In the EuroWordNet database, it is possible to 'project' these language-specific descendant word meanings to the other language (translate via the ILI). The result is a set of word meanings in the target language connected to the source language meanings via ILI-records. By taking the intersection of this projection in both directions we get an idea of the overlap of these semantic clusters (for further details, see (Peters et al., forthcoming)).

Table 70 gives the results of this mapping in both directions for each hierarchical node, once before the ILI-extension (rows headed by ILI-0) and once after the update (rows headed by ILI-1). For each language, the first column gives the total number of (sub)hyponyms per hierarchical node (the descendants), the second column gives the number of word meanings that have been projected to that particular language (from Spanish to Dutch and from Dutch to Spanish) and the third column lists the percentages of the projection for the total set of descendant word meanings.¹⁸ The last two columns give, for each language, the intersection of the projected word meanings (WMs in table above and the descendant word meanings, in absolute numbers and percentages. The bottom rows list the totals.

The general tendency for the Dutch wordnet is that the projection increased by 5.8%, whereas the increase of the intersection is 2.25%. For the Spanish wordnet these figures are 3.13% and 2.01% respectively. If we compare the increase of the projection (103 word meanings for Dutch and 56 for Spanish) with the increase in intersection (40 word meanings for Dutch and 36 for Spanish), we see that between 40-65% of the extended projection is effective, i.e. leads to an increase of the intersection. We suspect that the remaining incompatibilities either reflect a real difference in coverage or are caused by diverging classifications (e.g. *milk* is classified as a *product* instead of *comestible*; a hyponym of *food*).

¹⁷ In the case of animal and food, we have concentrated on the metonymic patterns. Because of the size of both sets, we have not investigated the instances of generalization.

¹⁸ In some cases, the projection extends the total set (more than 100%). This means that these words have been classified differently in the target language of the projection.

Table 71: Projection and Intersection increase Dutch-Spanish after adding sense-clusters to the ILI

		Projection to the Dutch WordNet					Projection to the Spanish WordNet				
		Desc. WMs	Projection		Intersection		Desc. WMs	Projection		Intersection	
			WMs	% of ES Desc.	WMs	% of NL Desc.		WMs	% of NL Desc.	WMs	% of ES Desc.
organiza- tion	ILI-0	48	47	97,92%	19	39,58%	186	41	22,04%	21	11,29%
	ILI-1	48	66	137,50%	20	41,67%	186	49	26,34%	26	13,98%
construc- tion	ILI-0	344	254	73,84%	131	38,08%	548	130	23,72%	77	14,05%
	ILI-1	344	270	78,49%	134	38,95%	548	139	25,36%	81	14,78%
food	ILI-0	154	133	86,36%	71	46,10%	533	83	15,57%	68	12,76%
	ILI-1	154	136	88,31%	71	46,10%	533	93	17,45%	69	12,95%
move	ILI-0	1183	445	37,62%	309	26,12%	384	392	102,08%	143	37,24%
	ILI-1	1183	510	43,11%	345	29,16%	384	418	108,85%	168	43,75%
sound	ILI-0	47	33	70,21%	18	38,30%	139	43	30,94%	19	13,67%
	ILI-1	47	33	70,21%	18	38,30%	139	46	33,09%	20	14,39%
Total	ILI-0	1776	912	51,35%	548	30,86%	1790	689	38,49%	328	18,32%
	ILI-1	1776	1015	57,15%	588	33,11%	1790	745	41,62%	364	20,34%
Increase			103	5,80%	40	2,25%		56	3,13%	36	2,01%

If we examine the figure in more detail, we see the following tendencies:

- the methodology is effective for *organization*, *construction* and *move*;
- the methodology is hardly effective for *food* and *sound*;

In the case of *move* (see table 70) we can expect that the effect is high because the extension (the composite ILIs) already makes up 25% of the total set of descendant senses. In the case of *organization* and *construction*, it is more remarkable because the extension only makes up 6-8% of the total of descendants. Further inspection shows that the effect for *construction* and *organization* is evenly spread over metonymy and generalization (50-70%) whereas the effect for *move* is almost exclusively due to generalization (97%). The fact that the effect is small for *sound* is in line with the low extension with composite ILI-records (6% of the total number of descendants). For *food*, the effect is more disappointing, given the much higher proportion of composite ILIs (11%).

To verify the quality of the extension, we have manually inspected the new word meanings that were projected from the Spanish wordnet to the Dutch wordnet. This inspection showed hardly any projections that are incompatible with the classifications of the projection before the extension, except for those that fall within the metonymic extension. In so far as there is a degree of variation in classification across the wordnets ((Peters et al., forthcoming), the extension is not worsening this effect. However, there is metonymic over-generation across the wordnets :

Table 72: Errors generated by automatically derived Composite ILIs

	New Projections to NL after the ILI Extension	Metonymic Overgeneration	Genuine Errors
food	3	1	0
construction	16	3	0
organization	19	4	0
move	65	0	4

Metonymic over-generation was to be expected, since regular polysemy does not necessarily hold across the languages. It may be caused by a cultural difference (e.g. not all *plants* and *animals* are considered to be *food* in all language/cultures), although we did not find any examples of this type of over-generation. Another possible reason for over-generation is a difference in lexicalization (e.g. metonymic meanings can be lexicalized by different word forms). In the case of *plant/food*, there is only one occurrence of over-generation: in the compound *vanilleplant* (the plant from which vanilla is extracted) the headword *plant* blocks the *spice* interpretation. The same phenomenon occurs more often with *organization/construction*, because a number of Dutch compounds can only refer to a building, such as *vestigingswerk* (defense construction), and *verenigingsgebouw* (building where the club is seated). Among the *constructions* we find several genuine cases of over-generation: *gemeenschap* (the community), *godsdiens*t (religion), *delegatie* (delegate), *commissie* (commission) are all groups of people without an associated building. Finally, in the

case of *move*, three errors occur: *bidden* (to pray), *gelijkspelen* (to finish a game with even scores) and *verschrijven* (to make a mistake with writing). However, these are due to incorrect translations or dubious classifications which also occur within non-extended projections. Metonymic over-generation is not problematic since it is up to the builder of the local wordnet to decide whether to include the metonymic pattern for a particular language. For example, in the cases discussed above the Dutch wordnet will only have an *eq_synonym* relation with one of the senses related by metonymy, while in other languages we may find the same word linked to multiple meanings.

5 Conclusions

In this document, we have described the work carried out for the first subset in EuroWordNet. The subset has been constructed starting from the common set of Base Concepts, including all the major synsets on which the other concepts depend. This set has been extended top-down by each site separately. The resulting wordnets have been described in terms of quantity (entries, senses, synsets, language-internal relations and equivalence relations), by comparing the overlap with Parole lexicons and by measuring the conceptual coverage by clustering of Top-Ontology concepts. Whereas the Spanish wordnet already has reached full coverage (advancing the planning), the Dutch wordnet has just covered the first subset with a higher density of language internal relations, and the Italian wordnet has full coverage but lacks equivalence relations. The distribution of the wordnets over the top-ontology was surprisingly balanced. Some slight imbalances for 1stOrder Entities have to be corrected. Similarly, the overlap with the top-frequent Parole entries is also very high. Missing entries can easily be added.

In addition we have carried out two comparisons to get an impression of the consistency of the wordnets. Both comparisons showed promising results. The in-depth comparison of 18 fields showed reasonable intersections. Most of the mistakes are due to translation errors. Alternative classifications can be used to encode multiple hyperonym. A similar conclusion has been made from the overall comparison. There is a high degree of overlap between subsequences and sequences with 1 gap. By filling these gaps we can improve the coverage in a coordinated way. Furthermore, extremely tangled graphs (Dutch verbs) are mostly due to generation of wrong translations.

The following improvements will therefore be made to the wordnets in the next building phase:

- improve balancing of 1stOrderClusters (Dutch and Italian)
- extend with missing top-frequent Parole entries (Dutch, Italian, Spanish)
- extend coverage (Dutch)
- check translations of extremely long hyponymy chains (especially Dutch verbs)
- fill sequences with 1 gap (Italian, Spanish and Dutch)
- extend translations (Italian)
- improve translation heuristics (Spanish and Dutch)

Finally, this deliverable describes the work done for updating the Inter-Lingual-Index (ILI) that interconnects the different wordnets. We showed that using fully automatic techniques we can achieve up to 5% improvement in matching across wordnets.

References

- Apresjan, J. (1973). *Regular Polysemy* In: Linguistics 142: 5-32
- Atkins, B (1993). *Building a Lexicon: The Contribution of Lexicography* In: International Journal of lexicography 3: 167-204
- Atserias J., S. Climent, J. Farreres, G. Rigau, H. Rodriguez, *Combining Multiple Methods for the Automatic Construction of Multilingual WordNets*, Proceedings of Conference on Recent Advances on NLP. RANLP 97. Tzgov Chark, Bulgaria, 1997.
- Buitelaar, P (1998). *Corelex: Systematic Polysemy and Underspecification*, Ph.D., Department of Computer Science, Brandeis University, Boston, U.S.A..
- Copestake A. and Briscoe, T (1991). *Lexical operations in a unification-based framework*. In: Pustejovsky J. and Bergler S. (eds.), Lexical Semantics and Knowledge Representation Association for Computational Linguistics.
- Copestake, A. (1995). *Representing Lexical Polysemy*, in: Proceedings of AAAI Stanford Spring Symposium, Stanford 1995
- Cruse, D.A. (1986). *Lexical Semantics*, Cambridge University Press, Cambridge, U.K. Dolan, W.B., (1994). *Word Sense Ambiguation: Clustering Related Senses*, COLING, Kyoto, Japan
- Dorr B., M.A. Marti & I. Castellon, *Spanish EuroWordNet and LCS-Based Interlingual Machine Translation*. Workshop on Interlinguas AMTA/SIG-IL. San Diego, US, 1997
- Evens, M.W. (ed.) (1988). *Relational Models of the Lexicon: Representing knowledge in semantic networks*, Cambridge, CUP.
- Fellbaum, C., Grabowski, J. & Landes, S., (1997). *Analysis of a Hand-Tagging Task*, In: Tagging Text with lexical Semantics: Why, What and How? SIGLEX workshop, Washington, U.S.A.
- Jorgensen, J., (1990). *The Psychological Reality of Word Senses*, In: Journal of Psycholinguistic Research vol. 19 no.3
- Kilgariff, A., (1997) *Evaluating Word Sense Disambiguation Programs: Progress Report*, In: Proceedings SALT workshop on Evaluation in Speech and Language Technology, Sheffield ,U.K
- Kilgariff, A., *I Don't Believe in Word Senses*, To appear in: Computers and the Humanities Special Issue on Word Sense Disambiguation
- Krovetz, R., (1996). *Homonymy and Polysemy in Information Retrieval*, In: Proceedings ACL97, Madrid, Spain
- Levin, B., (1993). *English Verb Classes and Alternations, a Preliminary Investigation*, University of Chicago Press, Chicago/London
- Miller G.A, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller, (1990). *Introduction to WordNet: An On-line Lexical Database*, In: International Journal of Lexicography, Vol. 3, No.4, 235-244.
- Nirenburg, S. (ed.), (1989). *Knowledge-based MT*, Special issue Machine Translation vol.4, no 1 and 2, Kluwer Publishers, Dordrecht, The Netherlands
- Nunberg, G & Zaenen, A., (1992). *Systematic Polysemy in Lexicology and Lexicography*, In: Proceedings of EURALEX'92 University of Tampere, Finland.
- Ostler, N. and Atkins, S., (1991). *Predictable Meaning Shift: some linguistic properties of lexical implication rules*. In: Pustejovsky J. and Bergler S. (eds.), Lexical Semantics and Knowledge Representation Association for Computational Linguistics.
- Peters, W., Vossen, P., Diez-Orzas, P., Adriaens, G., *Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index*, To appear in: Computers and the Humanities Special Issue on EuroWordNet
- Peters, I. and Peters, W., *Extracting Regular Polysemic Patterns in WordNet*. Technical Report, University of Sheffield, United Kingdom
- Pustejovsky, J. (1995). *The Generative Lexicon*, MIT Press, Cambridge MA, U.S.A.
- Rodriguez, H., S. Climent, P. Vossen, L. Bloksma; A. Roventini, F. Bertagna, A. Alonge, W. Peters, *The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology*. In: Computer and the Humanities, Special Issue on EuroWordNet.
- Vossen, P. (1998). *Introduction to EuroWordNet*, To appear in: Computers and the Humanities, Special Issue on EuroWordNet

Appendix I In-depth comparison of semantic clusters by different sites

Appendix Ia Comparing to the Dutch wordnet

General Comments

- **Mistakes:** most mistakes in the Dutch wordnet are due to wrong translations. It turns out that taking the best 3 translations generated by heuristics generates too many wrong translations. This will be adjusted to the best 2 translations. Only a few mistakes are due to wrong classifications.
- **Alternative classification:** in many cases parts (e.g. parts of buildings) are classified as subtypes of the wholes in the Reference wordnets: e.g. a *room* is both a type of *construction* and a part of a *construction*. This is a systematic difference with the Dutch wordnet where such parts are systematically classified as a type of part and related to the whole by a meronym relation
- **Coverage:** the coverage of the Dutch wordnet is less than the other wordnets. This is because the other wordnets have included larger proportions in Subset1.
- **Equivalence matching:** the Spanish wordnet has a direct matching of synsets with the ILI only using eq_synonym relations; the Dutch and Italian wordnet also have other types of equivalence relations.

First Order Entities

Building

Major Nodes, hyponyms and equivalence relations

	<i>Hyperonymic Synsets</i>	<i>No of Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_Hypo</i>	<i>EQ_Rest</i>
ES	{construcción-4}	548	548	548	0	0	0	0
IT	{costruzione-1} {edificio-1} {manifattura-1} {dimora-2}	194	7	5	2	1	0	0
NL	{bouwwerk-1}	344	223	39	188	15	0	1
WN15	{construction 4}	1220	1220	1220				

Projection of Reference wordnets to Dutch Source wordnet

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	548	367	181	276
IT	194	1	6	21
NL	344			
WN15	1210	920	290	364

Comparing projections for the Dutch wordnet

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	276	487	133	143	211
IT	21	350	15	6	329
NL	344	X	X	X	X
WN15	264	538	170	194	174
Union of Reference WNs		540	170	194	174

Errors in Dutch Source:

- wrong translations: 6
- wrong classifications: 8

Errors in Reference: 0

Alternative classifications:

- movable constructions
- parts of buildings
- institutions

Variant Projection of Unmatched ILI-records from Reference wordnets to Dutch Source Wordnet

	<i>Unmatched ILIs</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC-Source Projection</i>
Total		1990	1700	290	426	34

Comestibles

Major Nodes, hyponyms and equivalence relations

	<i>Hyperonymic Synsets</i>	<i>No of Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_ S</i>	<i>EQ_N S</i>	<i>EQ_Hyp er</i>	<i>EQ_Hy po</i>	<i>EQ_Re st</i>
ES	{alimento-1}	533	533					
IT	{cibo-1}	157	51	40	10	1		
NL	{voedsel-1}	151	154	27	129		1	
WN15	{food-1}	2123	2123					

Projection of Reference wordnets to Dutch Source wordnet

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	533	410	123	135
IT	51	30	21	35
WN15	2123	1923	200	191

Comparing projections for the Dutch wordnet

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	135	216	70	65	81
IT	35	162	24	11	127
WN15	191	239	103	88	48
Union of Reference WNs	195	243	103	92	48

Errors in Dutch Source:

- wrong translations: 6
- wrong classifications: 0

Errors in Reference: 0

Alternative classifications:

- natural products such as fruits, grain, corn, seeds
- drinks
- parts

Variant Projection of Unmatched ILI-records from Reference wordnets to Dutch Source Wordnet

	<i>Unmatched ILIs</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC- Source Projection</i>
Total		3199	3011	188	248	19

Container

Major Nodes, hyponyms and equivalence relations

	<i>Hyperonymic Synsets</i>	<i>No of Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_ S</i>	<i>EQ_N S</i>	<i>EQ_Hyp er</i>	<i>EQ_Hy po</i>	<i>EQ_Re st</i>
ES	{contenedor-2}	245	245					
IT	{contenitore-1}	161	7	6		1		
NL	{bak-1} {bergplaats-1}	26	31	11	22	2		
WN15	{container-1}	567	567					

Projection of Reference wordnets to Dutch Source wordnet

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	245	209	36	43
IT	7	2	5	12
WN15	567	505	62	57

Comparing projections for the Dutch wordnet

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	43	55	14	29	12
IT	12	30	8	4	18
WN15	57	68	15	42	11
Union of Reference WNs	59	70	15	44	11

Errors in Source:

- wrong translations: 13
- wrong classifications: 0

Errors in Reference: 7

Alternative classifications:

- voorwerp (object)

Variant Projection of Unmatched ILI-records from Reference wordnets to Dutch Source Wordnet

	<i>Unmatched ILIs</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC- Source Projection</i>
Total		1046	926	120	184	3

Covering

Major Nodes, hyponyms and equivalence relations

	<i>Hyperonymic Synsets</i>	<i>No of Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_ S</i>	<i>EQ_N S</i>	<i>EQ_Hyp er</i>	<i>EQ_Hy po</i>	<i>EQ_Re st</i>
ES	{cubierta-1} {cubierta-7}	425	425					
IT	{involucro-1} {copertura-2}	40	3	1	2			
NL	{bedekking-1}	27	43	3	29		2	
WN15	{covering-4} {covering-5}	1024	1024					

Projection of Reference wordnets to Dutch Source wordnet

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	425	330	95	113
IT	3	0	3	5
WN15	1024	876	148	147

Comparing projections for the Dutch wordnet

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	113	123	17	96	10
IT	5	27	5	-	22
WN15	147	156	18	129	9
Union of Reference WNs	147	156	18	129	9

Errors in Source:

- wrong translations: 31
- wrong classifications: 1

Errors in Reference: 7

Alternative classifications:

- garments
- parts of garments

Variant Projection of Unmatched ILI-records from Reference wordnets to Dutch Source Wordnet

	<i>Unmatched ILIs</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC- Source Projection</i>
Total		1846	1615	231	351	10

High Order Entities

Feeling

Major Nodes, hyponyms and equivalence relations

	<i>Hyperonymic Synsets</i>	<i>No of Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_Hypo</i>	<i>EQ_Rest</i>
ES	{sentimiento-1} {sensación-6} {sentir-3} V {sentir-5} V	253	253					
IT	{sentimento-1} {percepire-1} V {provare-7} V	178	32	24	9	2		
NL	{voelen-4} V {voelen-5} V {gevoel-2} {gevoel-3}	87	139	19	125			
WN15	{feeling-1} {feeling-1} {experience-6} V {feel-7} V {feel-8} V	448	448					

Projection of Reference wordnets to Dutch Source wordnet

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	253	212	411	56
IT	32	14	18	38
WN15	448	398	50	48

Comparing projections for the Dutch wordnet

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	56	126	17	39	70
IT	38	116	9	29	78
WN15	48	110	25	23	62
Union of Reference WNs	80	138	29	51	58

Errors in Source:

- wrong translations: 14
- wrong classifications: 3

Errors in Reference: 2

Alternative classifications:

- stimulus: aanvoelen (cause to feel like)
- experience: gewaarwording; ervaring; ervaren; meemaken; ondergaan; gewaarworden; waarnemen;
- attitude: houding; gemoedstoestand; bui/ stemming;
- ability: vermogen

Variant Projection of Unmatched ILI-records from Reference wordnets to Dutch Source Wordnet

	<i>Unmatched ILIs</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC-Source Projection</i>
Total		1620	1283	337	449	29

Phenomena

Major Nodes, hyponyms and equivalence relations

	<i>Hyperonymic Synsets</i>	<i>No of Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_ S</i>	<i>EQ_N S</i>	<i>EQ_Hyp er</i>	<i>EQ_Hy po</i>	<i>EQ_Re st</i>
ES	{ fenómeno-1 } { caer-57 }	415	415					
IT	{ fenomeno-1 }	100	23	19	7			
NL	{ verschijnsel-1 }	353	241	22	219	3	1	1
WN15	{ phenomenon-1 }	1012	1012					

Projection of Reference wordnets to Dutch Source wordnet

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	415	327	88	102
IT	23	5	18	27
WN15	1012	897	115	118

Comparing projections for the Dutch wordnet

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	102	445	10	92	343
IT	27	344	6	21	347
WN15	118	455	16	102	337
Union of Reference WNs	123	460	16	107	337

Errors in Source:

- wrong translations: 9
- wrong classifications: 2

Errors in Reference: 2

Alternative classifications:

- process/ change/ condition proces-2; verandering-1; gesteldheid-1 (all more general)
- systems: systeem (mechanisme)
- weather: weersgesteldheid (weather condition)
- power/force: energie-2 -> kracht-6 -> vermogen-; krachtveld
- possibilities: mogelijkheid
- diseases: ziekte-1

Variant Projection of Unmatched ILI-records from Reference wordnets to Dutch Source Wordnet

	<i>Unmatched ILIs</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC- Source Projection</i>
Total		2591	2119	472	790	17

Appendix Ib Comparing the Spanish wordnet

General Comments:

• Areas compared are:

- Garment, Places, Furniture, Plants (FirstOrder)
- Cooking, Sounds (Second Order)
- *Plants* can not be properly compared to WN1.5 due to the enormous difference of sets (WN1.5 contains a huge scientific taxonomy). Therefore in this area comparison is only performed using IT and NL as reference

• Major Nodes, hyponyms and equivalence relations (WN1.5 is not considered here)

- SpWN practically does not have eq relations other than eq_synonym —while NL and IT usually also have other types of equivalence
- SpWN correspondence to ILI is practically one-to-one, so number of SpWN synsets equals number of ILIs —while NL tends to have more ILIs than synsets; and IT less ILIs than synsets
- SpWN tends to have more synsets than the others —at this stage of the project

• Projections of Reference WNs to Spanish

- SpWN tends to already have covered most of the synsets projected from other WNs —except from WN1.5, obviously

• Comparing projections

- SpWN has few wrong classifications, but an amount of alternative classifications and cases of logical polysemy —to be resolved using complex ILIs or conflating word meanings
- Both alternative and wrong classifications in SpWN are a reflex of classifications in WN1.5; therefore when such a diagnostic occurs, it applies to both WNs
- Some cases of wrong translation are found in SpWN.

(Notice that SpWN Synsets including a wrong variant translation are not counted as proper *wrong translations*, except if this happens to be the only one variant in the synset, since it is considered that the synset has a correct Spanish-ILI correspondence —the refinement work to do in this case is to delete the weird variant.

E.g.: ILI-*car, auto* ↔ {coche, automóvil, *vagón} no comment
 ILI-*car, auto* ↔ {*vagón} wrong translation)

• Resolving missing synsets

- Missing synsets are not found using this procedure; the trend is to detect more cases of alternative classification

First Order Entities

Garment

Major Nodes, hyponyms and equivalence relations

	<i>Hyper Synsets</i>	<i>Nm. of hypos all levels</i>	<i>Nm. of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_hypo</i>	<i>EQ_rest</i>
ES	indumentaria-1,calzado-1	215	215	215	0	0	0	0
IT	indumento 1	156	3	3	0	0	0	0
NL	kledingstuk-1	23	36	5	32	0	0	0
WN15	wear-1, footwear-1, garment-1	277	272	277	0	0	0	0

Projection of Reference wordnets to Spanish Source wordnet

	<i>Reference hypos</i>	<i>No match in SpWN</i>	<i>Matching in SpWN</i>	<i>Synsets in SpWN</i>
ES	215			
IT	3	0	3	3
NL	36	10	26	26
WN15	277	149	128	128

Comparing projections for the Spanish wordnet

	<i>Projected SpWN Synsets</i>	<i>Union</i>	<i>Classification Intersection</i>	<i>Classification unique in Reference WN</i>	<i>Classification unique in SpWN</i>
ES	215	-	-	-	-
IT	3	215	3	0	212
NL	26	218	23	3	192
WN15	128	216	127	1	88
Union of Reference WNs	142	219	138	4	77

Errors in SpWN

Wrong translations: 12

(Some are genuine wrong translations, e.g. 'uplift' -a kind of bra- translated as 'construcción' -the act of building-; other are correct translations of the term in WN1.5, but the choice made by WN1.5 is extremely doubtful so SpWN inherits the error -e.g. WN15:'blue' as a dress, gloss 'she was wearing blue'; we translate automatically into 'azul')

Errors in Reference WNs

Wrong classifications: 4 (coverings which are not garment, e.g. screen in NL; 'wear' the act as the object in WN1.5; no possible comparison with IT)

Alternative Classifications: no

Variant Projection of Unmatched ILI-records from Reference wordnets to Spanish Source Wordnet

	<i>No Match in SpWN-Synsets</i>	<i>All Senses of ILIs</i>	<i>No Match in SpWN- All Senses</i>	<i>Matching ILI in SpWN</i>	<i>ILIs of Matching SpWN</i>	<i>Intersection with TC-SpWN Projection</i>
IT	0	0	0	0	0	0
NL	10	29	21	8	8	0
WN15	149	336	262	74	60	14
Total		349	273	76	62	14

No missing synsets in SpWN are found using this procedure

Furniture

Major Nodes, hyponyms and equivalence relations

	<i>Hyper Synsets</i>	<i>Nm. of hypos all levels</i>	<i>Nm. of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_hypo</i>	<i>EQ_rest</i>
ES	mobiliario 1	65	65	65	0	0	0	0
IT	mobile 2	75	4	4	0	0	0	0
NL	meubelstuk_1	11	15	9	6	0	0	0
WN15	furniture	174	174	174	0	0	0	0

Projection of Reference wordnets to Spanish Source wordnet

	<i>Reference hypos</i>	<i>No match in SpWN</i>	<i>Matching in SpWN</i>	<i>Synsets in SpWN</i>
ES	65			
IT	75	0	4	4
NL	15	6	9	9
WN15	174	109	65	65

Comparing projections for the Spanish wordnet

	<i>Projected Synsets</i>	<i>SpWN</i>	<i>Union</i>	<i>Classification Intersection</i>	<i>Classification unique Reference WN</i>	<i>in</i>	<i>Classification unique in SpWN</i>
ES	65						
IT	4		65	4	0		61
NL	9		69	5	4		60
WN15	65		65	65	0		0
Union of Reference WNs	69		69	65	4		0

Errors in SpWN: no

Errors in Reference WNs

Wrong translations: 1

Alternative Classifications: artifact (for 3 unique in NL, which are kinds of small cupboards)

Variant Projection of Unmatched ILI-records from Reference wordnets to Spanish Source Wordnet

	<i>No Match in SpWN- Synsets</i>	<i>All Senses of ILIs</i>	<i>No Match in SpWN- Senses</i>	<i>Matching ILI in SpWN</i>	<i>ILIs of Matching SpWM</i>	<i>Intersection with TC- SpWN Projection</i>
ES						
IT	0					
NL	6	41	26	15	13	2
WN15	109	238	193	45	39	6
Total		238	193	45	39	6

Some more alternative 'artifact' classifications detected by this procedure

Places

Major Nodes, hyponyms and equivalence relations

	<i>Hyper Synsets</i>	<i>Nm. of hypos all levels</i>	<i>Nm. of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_hypo</i>	<i>EQ_rest</i>
ES	lugar-1	373	373	373	0	0	0	0
IT	luogo1,luogo2	54	26	21	4	1	0	0
NL	plaats_1	533	424	95	346	12	4	1
WN15	location 1	1881	1881	1881	0	0	0	0

Projection of Reference wordnets to Spanish Source wordnet

	<i>Reference hypos</i>	<i>No match in SpWN</i>	<i>Matching in SpWN</i>	<i>Synsets in SpWN</i>
ES	373			
IT	26	1	25	25
NL	424	127	297	298
WN15	1881	1508	373	373

Comparing projections for the Spanish wordnet

	<i>Projected Synsets</i>	<i>SpWN</i>	<i>Union</i>	<i>Classification Intersection</i>	<i>Classification unique in Reference WN</i>	<i>Classification unique in SpWN</i>
ES	373					
IT	25		382	16	9	357
NL	298		613	58	240	315
WN15	373		373	373	0	0
Union of Reference WNs	619		619	373	246	0

Errors in SpWN

Wrong translations: 15

Wrong Classification: 1 (way/road as artifact)

Errors in Reference WNs

Wrong translations: 15

Wrong classification: 17 (mainly anatomical terminology: 'callosity', 'tuberosity' as places; also other as 'rubbish' as place; 'plant' as place; 'opening/gap' as place)

Doubtful classifications: container as a place - SpWN hasn't got the 245 hyponyms of 'container' classified as places, while others have -; rack/stand as a place

Alternative Classifications: cognition (for imaginary places); geographyc terms→land → object (e.g. 'depression', 'island', 'tundra', 'peninsula'); building → artifact (e.g. 'office'); facility/installation (e.g. sports fields)

Notice constructions and installations are clear cases of logical polysemy.

Variant Projection of Unmatched ILI-records from Reference wordnets to Spanish Source Wordnet

	<i>No Match in SpWN-Synsets</i>	<i>All Senses of ILIs</i>	<i>No Match in SpWN-Senses</i>	<i>Matching All ILI in SpWN</i>	<i>ILIs of Source WM</i>	<i>Matching Intersection with TC-SpWN Projection</i>
ES						
IT	1	33	15	18	13	5
NL	127	636	433	203	182	21
WN15	1508	2896	2423	473	422	51
Total		3303	2714	589	535	54

Many cases of alternative classifications noticed before (containers, some imaginary spaces,...) appear here as missing - not classified as places-in SpWN

Plants

Only first 3 levels of WN1.5 are used for comparison

Major Nodes, hyponyms and equivalence relations

	<i>Hyper Synsets</i>	<i>Nm. of hypos all levels</i>	<i>Nm. of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_hypo</i>	<i>EQ_rest</i>
ES	planta1	467	468	467	0	0	0	1
IT	pianta1	474	261	253	6	10	0	0
NL	plant1 gewas1	28	40	7	29	0	2	3
WN15	plant1 (first 3 levels)	802	802	802	0	0	0	0

Projection of Reference wordnets to Spanish Source wordnet

	<i>Reference hypos</i>	<i>No match in SpWN</i>	<i>Matching in SpWN</i>	<i>Synsets in SpWN</i>
ES	467			
IT	261	86	175	175
NL	40	7	33	33
WN15	802	670	132	132

Comparing projections for the Spanish wordnet

	<i>Projected SpWN Synsets</i>	<i>Union</i>	<i>Classification Intersection</i>	<i>Classification unique in Reference WN</i>	<i>Classification unique in SpWN</i>
ES					
IT	175	472	170	5	297
NL	33	481	19	14	448
WN15	132	467	132	0	335
Union of Reference WNs	272	486	253	19	214

Errors in SpWN: no

Errors in Reference WNs:

Wrong classification: 5

Wrong translation: 3

Alternative Classifications: microorganism (for 'alga'); vegetables (for edible roots and seeds)

Variant Projection of Unmatched ILI-records from Reference wordnets to Spanish Source Wordnet

	<i>No Match in SpWN-Synsets</i>	<i>All Senses of ILIs</i>	<i>No Match in SpWN- All Senses</i>	<i>Matching ILI in SpWN</i>	<i>ILIs of Matching SpWM</i>	<i>Intersection with TC-SpWN Projection</i>
ES						
IT	86	253	201	52	46	6
NL	7	37	24	13	9	4
WN15	670	925	870	55	42	13
Total		1158	1046	112	89	23

More cases of Alternative classification -Vegetables- found; also alternative classification: Fruits (kiwi, peanut), found
Wrong classification in SpWN found: Mistletoe as parasite but not as plant.

Higher Order entities

Sounds

Major Nodes, hyponyms and equivalence relations

	<i>Hyper Synsets</i>	<i>Nm. of hypos all levels</i>	<i>Nm. of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_hypo</i>	<i>EQ_rest</i>
ES	sonido-2, SONAR-3, emitir_sonidos-1	139	139	139	0	0	0	0
IT	{rumore 1, suono 1} {emettere 3, produrre 5}	45	41	21	21	2	0	0
NL	{geluid_2n}, {klinken_2v}	22	33	9	24	0	0	0
WN15	sound_13v, sound_5n, utter-3v	271	271	271	0	0	0	0

Projection of Reference wordnets to Spanish Source wordnet

	<i>Reference hypos</i>	<i>No match in SpWN</i>	<i>Matching in SpWN</i>	<i>Synsets in SpWN</i>
ES	139	-	-	-
IT	41	12	29	29
NL	33	5	28	28
WN15	271	132	139	139

Comparing projections for the Spanish wordnet

	<i>Projected SpWN Synsets</i>	<i>Union</i>	<i>Classification Intersection</i>	<i>Classification unique in Reference WN</i>	<i>Classification unique in SpWN</i>
ES	139				
IT	29	149	19	10	120
NL	28	151	16	12	123
WN15	139	139	139	0	0
Union of Reference WNs	160	160	139	21	0

Errors in SpWN

Wrong Classification: 2

Errors in Reference WNs

Wrong classification: 3

Wrong translation: 4

Alternative Classifications: communicate, breathe

Variant Projection of Unmatched ILI-records from Reference wordnets to Spanish Source Wordnet

	<i>No Match in SpWN-Synsets</i>	<i>All Senses of ILIs</i>	<i>No Match in SpWN-All Senses</i>	<i>Matching ILI in SpWN</i>	<i>ILIs of Matching SpWM</i>	<i>Intersection with TC-SpWN Projection</i>
ES						
IT	29	145	92	53	40	13
NL	28	52	38	14	12	2
WN15	139	682	415	267	208	59
Total		767	477	290	227	63

Another alternative classification detected: music

Cooking

Major Nodes, hyponyms and equivalence relations

	<i>Hyper Synsets</i>	<i>Nm. of hypos all levels</i>	<i>Nm. of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_hypo</i>	<i>EQ_rest</i>
ES	cocer-3v, cocina-1n	20	20	20	0	0	0	0
IT	cuocere 1v, preparare 3v, cuocere 2v	24	13	8	7	3	0	0
NL	klaarmaken_2v, koken_2v	3	6	0	6	0	0	0
WN15	cook1v, cook-2v, cook-3v, cook-4v, cooking-1n	57	57	57	0	0	0	0

Projection of Reference wordnets to Spanish Source wordnet

	<i>Reference hypos</i>	<i>No match in SpWN</i>	<i>Matching in SpWN</i>	<i>Synsets in SpWN</i>
ES	20			
IT	13	2	11	11
NL	6	2	4	4
WN15	57	34	23	23

Comparing projections for the Spanish wordnet

	<i>Projected SpWN Synsets</i>	<i>Union</i>	<i>Classification Intersection</i>	<i>Classification unique in Reference WN</i>	<i>Classification unique in SpWN</i>
ES	20				
IT	11	24	7	4	13
NL	4	23	1	3	19
WN15	23	23	20	3	0
Union of Reference WNs	26	26	20	6	0

Errors in SpWN: no

Errors in Reference WNs: no

Alternative Classifications: creation (for the act of cooking), change (for 'caramelize')

Variant Projection of Unmatched ILI-records from Reference wordnets to Spanish Source Wordnet

	<i>No Match in SpWN-Synsets</i>	<i>All Senses of ILIs</i>	<i>No Match in SpWN-All Senses</i>	<i>Matching ILI in SpWN</i>	<i>ILIs of Matching SpWM</i>	<i>Intersection with TC-SpWN Projection</i>
ES						
IT	2	3	3	0	0	0
NL	2	3	2	1	1	0
WN15	34	103	80	23	22	1
Total	34	103	80	23	23	1

Another Alternative Classification found: creation

Appendix Ic Comparing the Italian wordnet

First Order Entities

Animal

Hyponyms for “animal 1” with the type of equivalences

	<i>Hyperonymic Synsets</i>	<i>No Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_Hypo</i>	<i>EQ_Rest</i>
ES	{ animal-1 }	682	682	682	0	0	0	0
IT	{ animale 1, bestia 1, organismo animale 20 }	563	318	302	1	0	0	0
NL	{ dier_1, gedierte_1 }	26	43	13	23	0	9	0
WN15*	{ animal, animate being, beast, brute, creature, fauna }	2017	2017	2017	0	0	0	0

Projection of TC-Reference projection on Source wordnet

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	681	473	209	682
IT	563	0	318	318
NL	26	17	28	43
WN15*	2017	1794	224	2017
Total				

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	363	566	361	2	203
IT	495				
NL	157	565	156	1	408
WN15*	310	567	308	2	257
Union of Reference WNs					

Errors in source: 1

Variant Projection of Unmatched ILI-records from Reference wordnets to Source Wordnet

	<i>Unique Reference</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC-Source Projection</i>
ES	473	892	844	48	23	45
IT						
NL	15	13	10	3	2	1
WN15	1794*	2315	2246	69	62	38
Total						

* The comparison to WN1.5 has been performed on a smaller set of hyponyms, comprehensive of the first level below {animal, animate being, beast, brute, creature, fauna} and of the whole subsets of {bird} and {mammal}, via {chordate 1} and {vertebrate 1}.

Artist

Hyponyms for “**artist 1**” with the type of equivalences

	<i>Hyperonymic Synsets</i>	<i>No Hypos All Levels</i>	<i>No ILIs of</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_Hypo</i>	<i>EQ_Rest</i>
ES	{artista 2, pintor1}	30	30	30	0	0	0	0
IT	{artista 1}	91	38	33	7	3	0	0
NL	{kunstenaar 1} {artiest 1}	4	4	3	1	0	0	0
WN15	{artist 1}	71	71	71	0	0	0	0

Projection of TC-Reference projection on Source wordnet

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	30	25	6	30
IT				
NL	4	0	4	4
WN15	71	59	13	71
Total	105	84	23	105

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	6	96	2	4	92
IT					
NL	11	92	11	0	81
WN15	15	104	3	12	89
Union of Reference WNs	25	104	13	12	79

Errors in source: 0

Variant Projection of Unmatched ILI-records from Reference wordnets to Source Wordnet

	<i>Unique Reference</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC- Source Projection</i>
ES	25	58	53	5	39	56
IT						
NL	4	0	0	0	0	0
WN15	71	112	104	8	11	9
Total	100	170	157	13	50	65

Worker

Hyponyms for “**worker 2**” with the type of equivalences

	<i>Hyperonymic Synsets</i>	<i>No of Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_Hypo</i>	<i>EQ_Rest</i>
ES	trabajador 1	356	356	356	0	0	0	0
IT	lavoratore 1	552	251	204	55	29	0	0
NL	werknemer 1	9	11	5	7	0	0	0
WN15	worker 2	675	675	675	0	0	0	0

Projection of TC-Reference projection on Source wordnet

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	356	229	128	356
IT				
NL	9	4	7	11
WN15	675	523	153	675
Total				

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	344	653	262	82	291
IT					
NL	150	557	146	4	407
WN15	377	643	287	90	266
Union of Reference WNs		654	287	101	266

Errors in source: 2

Variant Projection of Unmatched ILI-records from Reference wordnets to Source Wordnet

	<i>Unique Reference</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC-Source Projection</i>
ES	229	444	403	41	95	16
IT						
NL	4	1	1	0	0	0
WN15	523	917	849	68	122	35
Total	756	1362	1253	109	217	51

Instrument

Hyponyms for “instrument 2” with the type of equivalences

	<i>Hyperonymic Synsets</i>	<i>No of Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_Hypo</i>	<i>EQ_Rest</i>
ES	{ herramienta-1, instrumento-3 }	185	185	185	0	0	0	0
IT	{ strumento 1, attrezzo 1, arnese 1, utensile 1 }	867	393	330	44	58	0	0
NL	{ instrument_1 }	437	266	72	199	31	1	0
WN15	{ instrument 2 }	509	509	509	0	0	0	0

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	185	101	85	186
IT	867	0	393	393
NL	437	163	103	266
WN15	509	379	131	509
Total				

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	163	1002	29	134	839
IT					
NL	180	1023	25	155	843
WN15					
Union of Reference WNs					

Alternative Classifications: for Italian, container has hyperonym Artifact, not Instrument

Variant Projection of Unmatched ILI-records from Reference wordnets to Source Wordnet

	<i>Unique Reference</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC-Source Projection</i>
ES	101	261	240	21	19	6
IT						
NL	163	344	322	22	23	2
WN15						
Total						

Vehicle

Hyponyms for “vehicle 1” with the type of equivalences

	<i>Hyperonymic Synsets</i>	<i>No of Hypos All Levels</i>	<i>No of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_Hyper</i>	<i>EQ_Hypo</i>	<i>EQ_Rest</i>
ES	{ transporte-5 }	189	189	189	0	0	0	0
IT	{ veicolo 1 }	172	72	60	6	15	0	0
NL	{ voertuig_1 }	21	35	7	27	1	0	0
WN15	{ vehicle 1 }	410	410	410	0	0	0	0

	<i>Reference Hyponyms</i>	<i>No Match in Source WordNet</i>	<i>Matching WMs in Source</i>	<i>Source WMs</i>
ES	189	134	56	190
IT	172	0	72	72
NL	21	23	12	35
WN15	410	343	68	410
Total				

	<i>Projected Source WMs</i>	<i>Union</i>	<i>Intersection</i>	<i>Unique in Reference WN</i>	<i>Unique in Source WN</i>
ES	97	176	94	3	79
IT					
NL	59	217	15	44	158
WN15	110	177	106	4	67
Union of Reference WNs					

Errors in source: *spartineve* is a mechanical device and not a vehicle: error of mapping.

Variant Projection of Unmatched ILI-records from Reference wordnets to Source Wordnet

	<i>Unique Reference</i>	<i>All Senses of ILIs</i>	<i>No Match in Source</i>	<i>Matching ILI in Source</i>	<i>ILIs of Matching Source WM</i>	<i>Intersection with TC-Source Projection</i>
ES	134	355	344	11	6	5
IT						
NL	23	68	66	2	4	0
WN15	343	681	658	23	15	11
Total						

Higher Order entities

Movement

Major Nodes, hyponyms and equivalence relations

	<i>Hyper Synsets</i>	<i>Nm. of hypos all levels</i>	<i>Nm. Of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_hyper</i>	<i>EQ_hypo</i>	<i>EQ_rest</i>
ES	movimiento 8, movimiento 2, movimiento 1, mover 1, mover 3, moverse 4	600	600	600	0	0	0	0
IT	movimento 1, muoversi 1, muovere 1	148	98	51	57	4	0	0
NL	beweging 1, bewegen 1, bewegen 2	1313	1304	94	1255	1	2	0
WN15	motion 1, motion 2, motion 5, move 1, move 4, move 2	1891						

Projections of TC-Reference on It WN

	<i>Reference hypos</i>	<i>No match in It WN</i>	<i>Matching in It WN</i>	<i>It WN WMs</i>
ES	600	553	47	69
IT	145	-	-	130
NL	1313	1200	104	142
WN15	1891	1817	74	102

Comparing projections

Comparing projections							
	<i>Projected Synsets</i>	<i>It</i>	<i>WN</i>	<i>Union</i>	<i>Classification Intersection</i>	<i>Classification unique Reference WN</i>	<i>Classification in unique in It WN</i>
ES	69			153	64	5	84
NL	142			200	90	52	58
WN15	102			155	95	7	53
Union of Reference WNs	164			203	109	55	39

Possibly wrong classification in It Wn: 4

Example errors in source: 1 (schema di gioco -play)

Alternative Classifications wrt the other Wns: 51. Examples alternative classifications:

- 'battuta' (-sport- the act of swinging or striking at a ball...) is a hyponym of 'azione' (act, action) via 'colpo' (the act of hitting);
- in Dutch some natural phenomena (like storm, shower ecc..) are hyponyms of 'movement', while, in Italian, they are 'atmospheric phenomena';
- 'play' (a preset plan of action) is classified as hyponym of 'movement';
- 'rabbriuidire' (feel shivers because of cold, fear, etc.) is classified as a perception in Italian.

Possible Missing Synsets

	<i>No Match in It WN- Synsets</i>	<i>All Senses of ILIs</i>	<i>No Match in It WN- All Senses</i>	<i>Matching ILI in It WN</i>	<i>ILIs of Matching It WM</i>	<i>Intersection with TC- It WN Projection</i>
ES	553	3925	3751	174	328	70
NL	1200	6878	6563	315	424	86
WN15	1817	7795	7438	357	592	89

Part of the synsets still need to be linked to ILI. Most problems seem however due to different classifications in the various Wns. A few cases can be reconsidered for a different classification in the It Wn.

Knowledge

Major Nodes, hyponyms and equivalence relations

	<i>Hyper Synsets</i>	<i>Nm. of all hypos levels</i>	<i>Nm. Of ILIs</i>	<i>EQ_S</i>	<i>EQ_NS</i>	<i>EQ_hyper</i>	<i>EQ_hypo</i>	<i>EQ_rest</i>
ES	Información 1, pensamiento 2, teoría 3, disciplina 2, pensamiento 1	159	159	159				
IT	Conoscenza 3, disciplina 1, conoscere 1	223	15	13	2			
NL	Kennis 2, Weten 2	53	69	20	53		1	

Projections of TC-Reference on It WN

	<i>Reference hypos</i>	<i>No match in It WN</i>	<i>Matching in It WN</i>	<i>It WN WMs</i>
ES	159	143	17	160
IT	223	-	-	
NL	53	53	16	51

Comparing projections

	<i>Projected It WN Synsets</i>	<i>Union</i>	<i>Classification Intersection</i>	<i>Classification unique in Reference WN</i>	<i>Classification unique in It WN</i>
ES	20	240		20	220
NL	51	235	5	11	219

Most of the Italian concepts still need to be linked to the ILI.

Alternative Classifications wrt the other Wns. Example alternative classifications: ‘Record’ (a document that can serve as legal evidence) in Dutch WordNet is a hyponym of information while in Italian Wordnet is a hyponym of textual matter.

Appendix II Software utilities for graph-comparison

We include here a list of programs that have been used to obtain all the statistical data for the overall comparison in this report. The list appears in alphabetical order. All this software can be found in the ftp-site of the project, as “tools/wns_compare.tar.gz”.

- **‘cadenes.pl’**: the aim of this program is to write to the standard output a list of the chain lengths with the number of occurrences of each length. This program needs in standard input a file of chains.

Syntax:

cadenes.pl < <in_file> > <out_file>

Example:

Query:

cadenes.pl < example.txt

Input:

```
00002728 00004865 00621770
00002728 00004865 02766721
00002728 00004865 03207851
00002728 00004865 05839075 06193747
00002728 00004865 05842570
00002728 00004865 05843454
00002728 00004865 05844200 05963844
```

Output:

```
3:      5
4:      2
```

- **‘chains.pl’**: the aim of this program is to query a file of chains created with the ‘*graf.pl*’ program which is described later. For all queries there are two consult modes, the single mode (only for counting occurrences) and the verbose mode (for counting and extracting occurrences). The chains have the next format:

```
<ili_record>(<english_ili?><spanish_ili?><dutch_ili?><italian_ili?>)
[<english_edge?><spanish_edge?><dutch_edge?><italian_edge?>]
<ili_record>(<english_ili?><spanish_ili?><dutch_ili?><italian_ili?>)
...
```

where the codes for describing coverage of nodes, (****), and edges, [****] consists of tuples of 4 elements, 1 if the corresponding language covers the node/edge or 0 in the other case.

The queries we can perform using this program are:

- **Complete Node Chains**: Giving a selection of languages, *L*, this query obtains the list of complete chains node-covered by all the languages in *L*.

Syntax:

chains.pl [-v] “cv(<languages>)” < <in_file> > <out_file>

where:

<languages> is: a sequence of the languages (using letters) we want to consult. For example: *ed* means *English* and *Dutch*.

[-v]: if this optional parameters is present, the program queries in verbose mode.

<in_file>: file containing the chains

<out_file>: file where the results will be placed

Example:

Query:

```
chains.pl -v "cv(s)" < example.chains
```

for getting all chains in *example.chains* completely node-covered for Spanish.

Input:

```
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05636823(1000){mental_energy} [1000] 05637150(1000){libidinal_energy} [1000]
05637285(1000){cathexis}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05636823(1000){mental_energy} [1000] 05636964(1100){incitement} [1100]
05637094(1100){goad}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05633277(1111){life}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636022(1100){hedonism}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1000]
05636402(1000){small_voice}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
05636540(1100){sense_of_duty}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
03839123(1100){superego}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1000] 05636665(1000){christ_within}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05634773(1000) [1000]{irrational_motive} 05635682(1000){compulsion}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
5634773(1000){irrational_motive} [1000] 05635349(1100){mania} [1000]
05635472(1000){monomania}
```

Output:

```
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05633277(1111){life}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636022(1100){hedonism}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
05636540(1100){sense_of_duty}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
03839123(1100){superego}
```

4 chains.

- **Complete Edge Chains:** Giving a selection of languages, *L*, this query obtains the list of complete chains edge-covered by all the languages in *L*.

Syntax:

```
chains.pl [-v] "ca(<languages>)" <in_file> > <out_file>
```

Example:

Query:

```
chains.pl -v "ca(d)" < example.chains
```

for getting all chains in *example.chains* completely edge-covered for Dutch.

Input:

```
00017008(0111){group} [0100] 05128528(0100){ethnic_group}
00017008(0111){group} [0100] 05144629(0100){subgroup}
00017008(0111){group} [0110] 05116306(0111){human_race}
```

Output:

```
00017008(0111){group} [0110] 05116306(0111){human_race}
1 chains.
```

- **Partial Node Chains:** Giving a selection of languages, L , and a threshold Min , this query obtains the list of chains, having a subchain of length not less than Min nodes, node-covered by all the languages in L .

Syntax:

chains.pl [-v] “**pv**(*<languages>*){*<subchain_length>*}” < *<in_file>* > *<out_file>*

where:

<subchain_length>: is a natural number. It defines the minimum length of the subsequence of nodes to search.

Example:

Query:

```
chains.pl -v “pv(di){2}” < example.chains
```

for getting all chains in *example.chains* containing subchains of length greater or equal to 2 node-covered for Dutch and Italian.

Input:

```
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05636823(1000){mental_energy} [1000] 05637150(1000){libidinal_energy} [1000]
05637285(1000){cathexis}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05636823(1000){mental_energy} [1000] 05636964(1100){incitement} [1100]
05637094(1100){goad}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05633277(1111){life}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636022(1100){hedonism}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1000]
05636402(1000){small_voice}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
05636540(1100){sense_of_duty}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
03839123(1100){superego}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1000] 05636665(1000){christ_within}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05634773(1000) [1000]{irrational_motive} 05635682(1000){compulsion}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05634773(1000){irrational_motive} [1000] 05635349(1100){mania} [1000]
05635472(1000){monomania}
```

Output:

```
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05633277(1111){life}
1 subchains.
```

- **Partial Edge Chains:** Giving a selection of languages, *L*, and a threshold *Min*, this query obtains the list of chains, having a subchain of length not less than *Min* edges, edge-covered by all the languages in *L*.

Syntax:

chains.pl [-v] “pa(<languages>){<subchain_length>}” <in_file> > <out_file>

Example:

Query:

chains.pl -v “pa(s){2}” < example.chains

for getting all chains in *example.chains* containing subchains of length greater or equal to 2 edge-covered for Spanish.

Input:

```
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05636823(1000){mental_energy} [1000] 05637150(1000){libidinal_energy} [1000]
05637285(1000){cathexis}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05636823(1000){mental_energy} [1000] 05636964(1100){incitement} [1100]
05637094(1100){goad}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05633277(1111){life}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636022(1100){hedonism}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1000]
05636402(1000){small_voice}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
05636540(1100){sense_of_duty}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
03839123(1100){superego}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1000] 05636665(1000){christ_within}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05634773(1000){irrational_motive} 05635682(1000){compulsion}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
05634773(1000){irrational_motive} [1000] 05635349(1100){mania} [1000]
05635472(1000){monomania}
```

Output:

```
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05633277(1111){life}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636022(1100){hedonism}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1000]
05636402(1000){small_voice}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
05636540(1100){sense_of_duty}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1100] 05636133(1100){conscience} [1100]
03839123(1100){superego}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
05635834(1100){ethical_motive} [1000] 05636665(1000){christ_within}
6 subchains.
```

- **Partial Node Chains with Gaps:** Giving a selection of languages, L , a threshold Min , and a number of gaps G , this query obtains the list of chains, having a subchain of length not less than Min nodes, containing G (node) gaps, node-covered by all the languages in L .

Syntax:

chains.pl [-v] “**pv**(*<languages>*){*<subchain_length>*}[*<gaps_length>*]**]**” < *<in_file>* > *<out_file>*

where:

<gaps_length>: is the number of nodes involved in the gaps of the subchain.

Example:

Query:

chains.pl -v “pv(s){3}[1]” < example.chains

for getting all chains in *example.chains* containing subchains of length greater or equal to 3, with one gap, node-covered for Spanish.

Input:

```
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
    05636823(1000){mental_energy} [1000] 05637150(1000){libidinal_energy} [1000]
    05637285(1000){cathexis}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
    05636823(1000){mental_energy} [1000]    05636964(1100){incitement} [1100]
    05637094(1100){goad}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
    05633277(1111){life}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
    05635834(1100){ethical_motive} [1100]    05636022(1100){hedonism}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
    05635834(1100){ethical_motive} [1100]    05636133(1100){conscience} [1000]
    05636402(1000){small_voice}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
    05635834(1100){ethical_motive} [1100]    05636133(1100){conscience} [1100]
    05636540(1100){sense_of_duty}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
    05635834(1100){ethical_motive} [1100]    05636133(1100){conscience} [1100]
    03839123(1100){superego}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1100]
    05635834(1100){ethical_motive} [1000]    05636665(1000){christ_within}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
    05634773(1000) [1000]{irrational_motive} 05635682(1000){compulsion}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
    5634773(1000){irrational_motive} [1000] 05635349(1100){mania} [1000]
    05635472(1000){monomania}
```

Output:

```
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
    05636823(1000){mental_energy} [1000]    05636964(1100){incitement} [1100]
    05637094(1100){goad}
00012517(1101){psychological_feature} [1100] 00013299(1111){motivation} [1000]
    5634773(1000){irrational_motive} [1000] 05635349(1100){mania} [1000]
    05635472(1000){monomania}
```

2 subchains.

- **Partial Edge Chains with Gaps:** Giving a selection of languages, L , a threshold Min , and a number of gaps G , this query obtains the list of chains, having a subchain of length not less than Min edges, containing G (edge) gaps, edge-covered by all the languages in L .

Syntax:

chains.pl [-v] “**pa**(*<languages>*){*<subchain_length>*}[*<gaps_length>*]**]**” < *<in_file>* > *<out_file>*

Example:

Query:

chains.pl -v "pa(d){3}[1]" < example.chains
for getting all chains in *example.chains* containing subchains of length greater or equal to 3, with one gap, edge-covered for Dutch.

Input:

```
02657448(0111){instrument} [0001] 02010561(0101){mechanism} [0001]
02473560(0101){engine} [0001] 01991412(0111){conveyance} [0011]
03235595(0111){craft} [0001] 02051671(0111){aircraft} [0001]
02061345(0001){amphibian}
02657448(0111){instrument} [0001] 02010561(0101){mechanism} [0001]
02473560(0101){engine} [0001] 01991412(0111){conveyance} [0011]
03235595(0111){craft} [0001] 02051671(0111){aircraft} [0111]
02054514(0111){aeroplane}
02657448(0111){instrument} [0001] 02010561(0101){mechanism} [0001]
02473560(0101){engine} [0001] 01991412(0111){conveyance} [0011]
03235595(0111){craft} [0001] 02051671(0111){aircraft} [0001]
02595197(0001){hang_glider}
02657448(0111){instrument} [0001] 02010561(0101){mechanism} [0001]
02473560(0101){engine} [0001] 01991412(0111){conveyance} [0011]
03235595(0111){craft} [0001] 02051671(0111){aircraft} [0101]
02106213(0101){airship}
```

Output:

```
02657448(0111) {instrument} [0001] 02010561(0101){mechanism} [0001]
02473560(0101){engine} [0001] 01991412(0111){conveyance} [0011]
03235595(0111){craft} [0001] 02051671(0111){aircraft} [0111]
02054514(0111){aeroplane}
```

1 subchains.

- ‘**chains2pairs.awk**’: this program gets the complete chains and writes all its edges.

Syntax:

chains2pairs.awk <in_file> > <out_file>

Example:

Query:

chains2pairs.awk example.chains > example.pairs

Input:

```
00002403 00004262
00002403 00004022 00682831
00002403 00004022 00005489
00002403 00004885
```

Output:

```
00002403 00004022
00002403 00004262
00002403 00004885
00004022 00005489
00004022 00682831
```

- ‘**cicles.pl**’: this program takes as input a list of pairs and shows all the cycles of the list.

Syntax:

cicles.pl <in_file> > <out_file>

Example:

Query:

cicles.pl example.pairs

Input:

```
00002403 00002909
00002403 00005260
01121367 01046072
01046072 01121367
01121367 01121367
```

Output:

```
01121367 01046072 01121367
```

- **'graf.pl'**: this program projects a set of wordnets over another and writes the projection to a file in order to perform posterior queries. The chains generated have the next line format:

```
<ili_record>(<english_ili?><spanish_ili?><dutch_ili?><italian_ili?>)
[<english_edge?><spanish_edge?><dutch_edge?><italian_edge?>]
<ili_record>(<english_ili?><spanish_ili?><dutch_ili?><italian_ili?>)
```

...

Syntax:

graf.pl <base_wn> [<projected_wn> ...]

where:

<base_wn>: is the first letter of the language of the wordnet skeleton we want to load. This letter can be: *e* for English, *s* for Spanish, *d* for Dutch and *i* for Italian.

<projected_wn>: is the first letter of the language of the wordnet we want to project. This letter can be: *e* for English, *s* for Spanish, *d* for Dutch and *i* for Italian.

Example:

Query:

```
graf.pl d
```

Input: (we suppose that dutch wordnet is only the next pairs for this example).

```
00002403 00002909
00002403 00005260
00005260 00513550
00005260 00739927
00005260 01219174
```

Output:

```
00002403(0010) [0010] 00002909(0010)
00002403(0010) [0010] 00005260(0010) [0010] 00513550(0010)
00002403(0010) [0010] 00005260(0010) [0010] 00739927(0010)
00002403(0010) [0010] 00005260(0010) [0010] 01219174(0010)
```

- **'inclusions.awk'**: this program deletes all the chains not finishing in a leaf.

Syntax:

inclusions.awk <in_file> > <out_file>

Example:

Query:

```
inclusions.awk example.chains > cp.chains
```

Input:

```
00013338 01472320
00013338 01472320 01257491
00013338 01472320 01257491 00086015
00013338 01472320 01257491
00013338 01472320 01257491 00202465
```

Output:

```
00013338 01472320 01257491 00086015
00013338 01472320 01257491 00202465
```